A variational method for parameter estimation in a logistic spatial regression

Cécile Hardouin

Version of Record: https://www.sciencedirect.com/science/article/pii/S2211675318301854
Manuscript_b3605ec04b3dcb1d638ff8db8f53c95b

A Variational Method for Parameter Estimation in a Logistic Spatial Regression

Cécile Hardouin[a]

[a]*MODAL'X, Université Paris Nanterre, France*

Abstract

We consider a logistic regression. The spatial dependence is captured through a hidden Gaussian process after the logit transformation of the Bernoulli success probabilities. In a hierarchical framework, likelihood-based estimation requires an EM algorithm. However, the expectations in the E-step are not available in closed-from expressions. We propose a variational approximation of the complete likelihood, that has a Gaussian form. We then obtain the desired approximations of the expectations. We conduct a simulation study to compare our approach with Laplace approximation.

Keywords: logistic regression, variational estimation methods, EM algorithm, Laplace approximation

1. Introduction, motivation

Binary spatial data occur in various domains; in ecology or epidemiology, binary variables indicate absence or presence of a certain plant, or animal, or illness, on a two-dimensional domain. In economics and social sciences, binary data can be used for instance in contexts of standard adoption, voting models; in these contexts, the spatial feature is translated into a neighbourhood graph between agents. Binary data also occur in image analysis, for instance in texture analysis. More generally, one can also transform continuous data into binary responses, where we consider 1 (resp. 0) over (resp. under) a predefined threshold. We consider in this paper the logistic regression model. This model is well-known and used in many contexts, it allows to account for both spatial dependence and for the effects of potential covariates. Spatial logistic regression has been widely used for modeling land-use change; see for instance Tayyebi et al. (2010) and the recent works of Schneider and Pontius (2001) for deforestation analysis, Serneels and Lambin (2001) in agriculture (Serneels and Lambin, 2001), and Nong and Du (2011) for urban growth modeling. In these models, socio-economic and environmental variables are used as covariates while urban and non-urban areas are considered as binary outputs. Logistic regression is also widely used in various other contexts, for instance for cloud-covering (Wu and Zhang (2013), Sengupta et al. (2016)), or in disease mapping (Diggle and Giorgi (2015)).

Intrisically, inference involves a hidden unobserved process; then likelihood-based estimation procedures rely on the so-called completed likelihood, together with an EM algorithm (Dempster

Email address: hardouin@parisnanterre.fr (Cécile Hardouin)

et al. (1977)). However, the expectations in the E-step of the algorithm are not available in closed-form expressions. There are several ways to overcome this issue, see Paciorek (2007) for a review of the general techniques. A common approach is to use Monte Carlo procedures; see for instance Robert and Casella (2004), Cappe et al. (2005). Another solution is to use Laplace approximation to approximate the intractable integrals, see e.g. Spiegelhalter (1990) or Sengupta and Cressie (2013). We propose in this work an alternative method, using a deterministic approximation for the unknown conditional distribution of the hidden process given the observations. Our approach is known as a variational method. Variational methods have been used in physics, but they also appeared in machine learning context and more recently in statistics for estimation problems (see e.g. Rustagi (1976), Jaakola and Jordan (2000)). The key feature is to consider a lower bound on the complete likelihood, and optimize this lower bound. In this work, we consider a lower bound of the logistic function; setting this bound in the completed likelihood expression, we obtain a variationally transformed likelihood, which is our new objective function. This operation introduces supplementary parameters known as variational parameters. Consequently, our transformed likelihood involves both model parameters and variational parameters, but the main interest is that it has a Gaussian form, for fixed values of the variational parameters. Thus we obtain the expectations required in the E-step of the EM algorithm in closed-form expressions. Hence we can run the M-step to find the estimates of the model parameters. Then in turn, we update the variational parameters by an optimization procedure, the model parameters being fixed to the latest estimates. In summary, each iteration of the so-called Variational EM (VEM) algorithm is achieved in three steps, computation of the expectations, maximisation of the model parameters, and adjustment of the variational parameters.

The method can be compared to the Laplace approximation, which also utilizes a Gaussian approximation; particularly, the Laplace approximation needs to compute the mode of the objective function at each iteration of the EM algorithm, while the variational approximation involves extra parameters that need to be updated at each iteration. We conduct simulation experiments, running the two procedures, in order to evaluate the advantages and drawbacks of the methods.

The plan of this paper is as follows. In Section 2, we describe our process model for binary data, based on a hidden spatial Gaussian process model. Section 3 is devoted to parameter estimation using the variational approach; we present the variationally transformed likelihood and describe the Variational EM (VEM) algorithm for obtaining estimators. We conduct a simulation study in Section 4; we compute the estimates obtained from both Variational EM and ordinary EM with Laplace approximations, and compare the results of the two methods. We also investigate the properties of the estimators; first, since there is no theoretical result about the variance of variational estimators, we compute an approximation of the variance through a bootstrap approach. Then we study the large sample properties, conducting experiments for increasing lattice sizes. Finally we investigate how sensitive to the initial values of the algorithm the estimates are. We apply our VEM algorithm to a real data set in Section 5, and present a full procedure to propose initial values of the algorithm, ending by final variational estimates and their bootstrap variances. Conclusion follows in Section 6.

2. The process model

We consider a finite two-dimensional domain $D \equiv \{\mathbf{s}_i : i = 1, \ldots, n\} \subset \mathbf{R}^d$, with $\mathbf{s}_i = (s_{i1}, s_{i2})$ for $i = 1, ..., n$. Let $\mathbf{Z} = (Z(\mathbf{s}_1), ..., Z(\mathbf{s}_n))^{\mathrm{T}}$ be the process on D, taking its values in the state space $E = \{0, 1\}^n$. In a hierarchical framework, we model the variables $Z(.)$ as Bernoulli variables,

whose means depend on an underlying spatial process $\mathbf{Y} = (Y(\mathbf{s}_1), ..., Y(\mathbf{s}_n))^{\mathrm{T}}$. Moreover, we assume that these Bernoulli variables are conditionally independent, given the hidden process $\mathbf{Y}$.

Thus, for each $\mathbf{s} \in D$, we write the following independent conditional distributions for $Z(\mathbf{s})$ given $\mathbf{Y}$ as,

$$Z(\mathbf{s}) \mid Y(\mathbf{s}) \sim Ber(p(\mathbf{s})), \tag{1}$$

where

$$p(\mathbf{s}) = \frac{e^{Y(\mathbf{s})}}{1 + e^{Y(\mathbf{s})}}. \tag{2}$$

Then $P(Z(\mathbf{s}) = z \mid Y(\mathbf{s})) = p(\mathbf{s})^z(1 - p(\mathbf{s}))^{1-z} = \dfrac{1}{1 + e^{-Y(\mathbf{s})(2z-1)}}$.

Then we model the hidden process $\mathbf{Y}$ as the sum of two components:

$$Y(\mathbf{s}) = \mathbf{X}(\mathbf{s})^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon(\mathbf{s}). \tag{3}$$

The first term represents the large-scale spatial variation, or the trend; it is modeled as a linear combination of p known covariates $\mathbf{X}(\mathbf{s}) = (X_1(\mathbf{s}), \cdots, X_p(\mathbf{s}))^{\mathrm{T}}$, and $\boldsymbol{\beta}$ denotes the $p-$dimensional vector of the unknown regression coefficients. The second term holds for small-scale spatial variation, and we consider a zero-mean Gaussian spatial process $\boldsymbol{\varepsilon}$,

$$\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \boldsymbol{\Sigma}), \tag{4}$$

with unknown spatial covariance matrix $\boldsymbol{\Sigma}$. Thus, the model parameters that need to be estimated are $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. If we set a parametric assumption for $\boldsymbol{\Sigma}$, that is $\boldsymbol{\Sigma} = \sigma^2\mathbf{Q}(\boldsymbol{\theta})$, the full model parameters are thus $\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}$. We now present the parameters estimation procedure in the next section.

3. Parameter estimation

Let us note the parameters to be estimated $\boldsymbol{\varphi} = (\boldsymbol{\beta}, \boldsymbol{\Sigma})$, and let us take the notation $[U \mid V]$ for the conditional distribution of U given V. Since our hierarchical framework involves hidden process, we consider the complete likelihood instead of the true likelihood. However, we do not consider fully Bayesian inference. We do have a hierarchical model, but we do not put prior distributions on the parameters.

Let us write the complete log likelihood, L_c, for the unknown parameters, given the data. The complete data involves the observations $\mathbf{Z}$ and the unobserved $\boldsymbol{\varepsilon}$. Since we have the following decomposition,

$$[\mathbf{Z}, \boldsymbol{\varepsilon} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}] = [\mathbf{Z} \mid \boldsymbol{\varepsilon}, \boldsymbol{\beta}] \times [\boldsymbol{\varepsilon} \mid \boldsymbol{\Sigma}], \tag{5}$$

we write the complete log likelihood as,

$$L_c(\mathbf{Z}, \boldsymbol{\varepsilon} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \ln[\mathbf{Z} \mid \boldsymbol{\varepsilon}, \boldsymbol{\beta}] + \ln[\boldsymbol{\varepsilon} \mid \boldsymbol{\Sigma}] \tag{6}$$

$$= -\sum_{\mathbf{s}\in D} \ln(1 + e^{Y(\mathbf{s})}) + \sum_{\mathbf{s}\in D} Y(\mathbf{s})Z(\mathbf{s}) - \frac{1}{2}\ln(\det \boldsymbol{\Sigma}) - \frac{1}{2}\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon} - \frac{n}{2}\ln 2\pi \tag{7}$$

where we recall $Y(\mathbf{s}) = \mathbf{X}(\mathbf{s})^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon(\mathbf{s})$. Our goal is to obtain maximum likelihood estimates of $\boldsymbol{\varphi} = (\boldsymbol{\beta}, \boldsymbol{\Sigma})$ maximizing (7). The process $\boldsymbol{\varepsilon}$ being not observed, estimation has to be performed using the EM algorithm, see Dempster et al. (1977), McLachlan and Krishnan (2008).

Let us define

$$q(\boldsymbol{\varphi}, \hat{\boldsymbol{\varphi}}^{(l)}) = E\left[L_c(\mathbf{Z}, \boldsymbol{\varepsilon} \mid \boldsymbol{\varphi}) \mid \mathbf{Z}, \hat{\boldsymbol{\varphi}}^{(l)}\right]. \tag{8}$$

Starting with an initialization $\hat{\boldsymbol{\varphi}}^{(0)}$, the $l-$th run of the algorithm is achieved in two steps:
The E (expectation) step is to compute $q(\boldsymbol{\varphi}, \hat{\boldsymbol{\varphi}}^{(l-1)})$.
The M (maximisation) step is maximizing $q(\boldsymbol{\varphi}, \hat{\boldsymbol{\varphi}}^{(l-1)})$ in order to obtain $\hat{\boldsymbol{\varphi}}^{(l)} = \arg\max_{\boldsymbol{\varphi}} q(\boldsymbol{\varphi}, \hat{\boldsymbol{\varphi}}^{(l-1)})$.

In our case the E-step is an issue since we do not know the conditional distribution of $\boldsymbol{\varepsilon}$ given the observations $\mathbf{Z}$. There are different ways to overcome this issue. One approach may be to approximate the expectations using Monte Carlo integration, and run a so-called stochastic EM (SEM) algorithm (see e.g. Robert and Casella (2004), McLachlan and Krishnan (2008)). The issue in this approach lies in the simulation, where a Metropolis algorithm is typically used to simulate $\boldsymbol{\varepsilon}$. Choosing the "right"proposal density (see Chib and Greenberg (1995), Roberts and Rosenthal (2001)) can be problematic, and computations can be very slow for large data sets. Another classical remedy is to apply self-normalized importance sampling (see Robert and Casella (2004), Section 3.3). In this case, choosing the "right"importance distribution can also be problematic; moreover, we can observe a degeneracy of the weights for large n, leading to poor estimates. Investigating closer, the main issue in both methods comes from the first term of the complete likelihood, $\sum_{\mathbf{s}\in D} \ln(1 + e^{Y(\mathbf{s})})$, which is directly derived from the logit function and hence the logistic regression model. Our alternative method replaces this term by another one which is no more problematic. A third method which is widely used is to proceed with Laplace approximations (see e.g. Sengupta and Cressie (2013)); they are based on second-order Taylor-series expansions of the logarithm of the integrands around their respective modes. Then the density of $\boldsymbol{\varepsilon}$ given the data and $\hat{\boldsymbol{\varphi}}^{(l)}$ is approximatively proportional to a Gaussian density; the method also allows to treat the problematic term $\sum_{\mathbf{s}\in D} \ln(1 + e^{Y(\mathbf{s})})$. Here we propose a variational method derived from an initial approximation of the logistic function, that we present below. Our method can be compared to the Laplace approximation which also uses a Gaussian approximation, but it is advantageous because it does not need to compute the mode at each iteration of the EM algorithm; the use of variational parameters offers larger flexibility, and the method allows for accurate approximation.

Roughly speaking, variational techniques are based on some approximation using extra parameters called variational parameters. Quoting Jaakola and Jordan (2000), for fixed values of the variational parameters, the transformed problem often has a closed form solution, providing an approximate solution to the original problem. Unfortunately, since variational estimates are based on an approximation of the true log likelihood, we don't have theoretical results on their consistency, or asymptotic normality. There's no general theory about variational estimators' properties, see e.g. Peyrard et al. (2018), paragraph 6.3. However, they are known to be empirically accurate. In the framework of binary data, our variational approach is based on an approximation of the logistic function, which was introduced by Jaakola and Jordan (2000). In a Bayesian and non spatial context, Jaakola and Jordan (2000) study a logistic regression model with a Gaussian prior on the parameter vector; they show that the approximate of the conditional posterior distribution contains the true conditional distribution. We develop their approach hereafter, in the framework of an added spatial Gaussian process.

Let us note the logistic function,

$$g(x) = \frac{e^x}{1 + e^x} = \frac{1}{1 + e^{-x}},$$

defined for any real x. Jaakola and Jordan (Jaakola and Jordan (2000)) give the following inequality for this function:

$$\ln g(x) \geq \ln g(\tau) + \frac{x-\tau}{2} - \lambda(\tau)(x^2 - \tau^2) \tag{9}$$

where $\lambda(\tau) = \frac{1}{4\tau}\tanh(\tau/2) = \frac{g(\tau) - 1/2}{2\tau}$. Moreover this lower bound is exact whenever $\tau^2 = x^2$. We apply this inequality to $-\ln(1 + e^{Y(\mathbf{s})}) = \ln g(-Y(\mathbf{s}))$, for each $\mathbf{s} \in D$. Let us note $\boldsymbol{\tau} = (\tau(\mathbf{s}_1), ..., \tau(\mathbf{s}_n))^{\mathrm{T}}$, we obtain

$$-\sum_{\mathbf{s}\in D} \ln(1 + e^{Y(\mathbf{s})}) + \sum_{\mathbf{s}\in D} Y(\mathbf{s})Z(\mathbf{s}) \geq T_1(\boldsymbol{\tau}) + T_2(\boldsymbol{\tau}, \boldsymbol{\beta}) + T_3(\boldsymbol{\tau}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$$

where

$$T_1(\boldsymbol{\tau}) = \sum_{\mathbf{s}\in D} \left\{ \ln g(\tau(\mathbf{s})) - \frac{\tau(\mathbf{s})}{2} + \tau(\mathbf{s})^2 \lambda(\tau(\mathbf{s})) \right\}, \tag{10}$$

$$T_2(\boldsymbol{\tau}, \boldsymbol{\beta}) = \sum_{\mathbf{s}\in D} \left\{ -\lambda(\tau(\mathbf{s}))(\mathbf{X}(\mathbf{s})^{\mathrm{T}}\boldsymbol{\beta})^2 + (\mathbf{X}(\mathbf{s})^{\mathrm{T}}\boldsymbol{\beta})(Z(\mathbf{s}) - \frac{1}{2}) \right\}, \tag{11}$$

and

$$T_3(\boldsymbol{\tau}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \sum_{\mathbf{s}\in D} \left\{ -\varepsilon(\mathbf{s})^2 \lambda(\tau(\mathbf{s})) + \varepsilon(\mathbf{s})[Z(\mathbf{s}) - \frac{1}{2} - 2\lambda(\tau(\mathbf{s}))(\mathbf{X}(\mathbf{s})^{\mathrm{T}}\boldsymbol{\beta})] \right\}. \tag{12}$$

That is, we have a lower bound for the complete log likelihood,

$$L_c(\mathbf{Z}, \boldsymbol{\varepsilon} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}) \geq \widetilde{L}_c(\mathbf{Z}, \boldsymbol{\varepsilon} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\tau}),$$

with

$$\widetilde{L}_c(\mathbf{Z}, \boldsymbol{\varepsilon} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\tau}) = T_1(\boldsymbol{\tau}) + T_2(\boldsymbol{\tau}, \boldsymbol{\beta}) + T_3(\boldsymbol{\tau}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) - \frac{1}{2}\boldsymbol{\varepsilon}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon} - \frac{1}{2}\ln(\det \boldsymbol{\Sigma}) + \text{const.} \tag{13}$$

Let us note that the problematic term $\sum \ln(1 + e^{Y(\mathbf{s})})$ is absent in this expression. Our new target is the variational lower bound $\widetilde{L}_c(\mathbf{Z}, \boldsymbol{\varepsilon} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\tau})$ defined in (13), which involves the model parameters and the so-called variational parameters $\boldsymbol{\tau}$. Moreover, the variational lower bound is exact for a particular choice of $\boldsymbol{\tau}$, which is $\tau(\mathbf{s})^2 = Y(\mathbf{s})^2$, for all $\mathbf{s} \in D$.

Starting with this initial choice, we have

$$L_c^{(0)}(\mathbf{Z}, \boldsymbol{\varepsilon} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \widetilde{L}_c^{(0)}(\mathbf{Z}, \boldsymbol{\varepsilon} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\tau}).$$

Then we alternately maximise $\widetilde{L}_c$ with respect to the model parameters, and update the variational parameters; we first search for $(\boldsymbol{\beta}_{\max}, \boldsymbol{\Sigma}_{\max})$ maximising $\widetilde{L}_c(\mathbf{Z}, \boldsymbol{\varepsilon} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\tau})$ for fixed $\boldsymbol{\tau}$, then the updated variational parameters are obtained maximising $\widetilde{L}_c(\mathbf{Z}, \boldsymbol{\varepsilon} \mid \boldsymbol{\beta}_{\max}, \boldsymbol{\Sigma}_{\max}, \boldsymbol{\tau})$ in $\boldsymbol{\tau}$. This leads to the following inequalities,

$$\widetilde{L}_c(\mathbf{Z}, \boldsymbol{\varepsilon} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\tau}) \leq \widetilde{L}_c(\mathbf{Z}, \boldsymbol{\varepsilon} \mid \boldsymbol{\beta}_{\max}, \boldsymbol{\Sigma}_{\max}, \boldsymbol{\tau}) \leq \widetilde{L}_c(\mathbf{Z}, \boldsymbol{\varepsilon} \mid \boldsymbol{\beta}_{\max}, \boldsymbol{\Sigma}_{\max}, \boldsymbol{\tau}_{\max}).$$

Our goal is to iterate this maximisation-update procedure in order to obtain $\widetilde{L}_c(\mathbf{Z}, \boldsymbol{\varepsilon} \mid \boldsymbol{\beta}_{\max}, \boldsymbol{\Sigma}_{\max}, \boldsymbol{\tau}_{\max}) \simeq L_c(\mathbf{Z}, \boldsymbol{\varepsilon} \mid \boldsymbol{\beta}_{\max}, \boldsymbol{\Sigma}_{\max})$ at the end.

As $L_c(\boldsymbol{Z}, \boldsymbol{\varepsilon} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\tau})$, the expression of $\widetilde{L}_c(\boldsymbol{Z}, \boldsymbol{\varepsilon} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\tau})$ involves unobserved variables and, classically, we run an EM algorithm (with an additional updating step of the variational parameters). The advantage of considering this variational transformation is that we obtain the desired expectations in closed-form expressions, as we now demonstrate. Indeed, we show that the conditional distribution $[\boldsymbol{Z}, \boldsymbol{\varepsilon} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\tau}]$ is proportional to a multivariate Gaussian distribution, for fixed variational parameter $\boldsymbol{\tau}$. Let us note $\mathbf{M} = (M(\mathbf{s}_1), ..., M(\mathbf{s}_n))^{\mathrm{T}}$, with

$$M(\mathbf{s}) = Z(\mathbf{s}) - \frac{1}{2} - 2\lambda(\tau(\mathbf{s}))(\mathbf{X}(\mathbf{s})^{\mathrm{T}}\boldsymbol{\beta}). \tag{14}$$

Then we write $T_3(\boldsymbol{\tau}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) - \frac{1}{2}\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{M} - \frac{1}{2}\boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{W}^{-1}\boldsymbol{\varepsilon}$, with

$$\mathbf{W}^{-1} = \boldsymbol{\Sigma}^{-1} + 2\boldsymbol{\Lambda}(\boldsymbol{\tau}), \tag{15}$$

where $\boldsymbol{\Lambda}(\boldsymbol{\tau})$ is a diagonal matrix with diagonal elements $\lambda(\tau(\mathbf{s}))$. We obtain,

$$\widetilde{L}_c(\boldsymbol{Z}, \boldsymbol{\varepsilon} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\tau}) = T_1(\boldsymbol{\tau}) + T_2(\boldsymbol{\tau}, \boldsymbol{\beta}) + \boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{M} - \frac{1}{2}\boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{W}^{-1}\boldsymbol{\varepsilon} - \frac{1}{2}\ln(\det\boldsymbol{\Sigma}) + \text{const.}$$

For fixed $\boldsymbol{\tau}$, the conditional distribution $[\boldsymbol{\varepsilon} \mid \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\tau}]$ is unknown, but it is proportional to $[\mathbf{Z}, \boldsymbol{\varepsilon} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\tau}]$. Denoting $\boldsymbol{\mu} = \mathbf{W}\mathbf{M}$, we write

$$p(\boldsymbol{\varepsilon} \mid \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\tau}) \propto \exp\left\{T_1(\boldsymbol{\tau}) + T_2(\boldsymbol{\tau}, \boldsymbol{\beta}) + \frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\mathbf{W}^{-1}\boldsymbol{\mu}\right\} \frac{1}{\sqrt{\det\boldsymbol{\Sigma}}} \exp\left\{-\frac{1}{2}(\boldsymbol{\varepsilon} - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{W}^{-1}(\boldsymbol{\varepsilon} - \boldsymbol{\mu})\right\}. \tag{16}$$

Moreover, evaluating the proportionality constant on the right-hand side of 16 yields:

$$[\boldsymbol{\varepsilon} \mid \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\tau}] = N(\boldsymbol{\mu}, \mathbf{W}) \tag{17}$$

Finally, our variational EM algorithm is based on the following expectation,

$$q(\boldsymbol{\varphi}, \hat{\boldsymbol{\varphi}}^{(l)}; \boldsymbol{\tau}) = E\left[\widetilde{L}_c(\mathbf{Z}, \boldsymbol{\varepsilon} \mid \boldsymbol{\varphi}, \boldsymbol{\tau}) \mid \mathbf{Z}, \hat{\boldsymbol{\varphi}}^{(l)}\right], \tag{18}$$

where the expectation is taken with respect to the conditional distribution (17) above. We deduce that,

$$q(\boldsymbol{\varphi}, \hat{\boldsymbol{\varphi}}^{(l)}; \boldsymbol{\tau}) = T_1(\boldsymbol{\tau}) + T_2(\boldsymbol{\tau}, \boldsymbol{\beta}) + \hat{\boldsymbol{\mu}}^{(l)\mathrm{T}}\mathbf{M} \tag{19}$$

$$- \frac{1}{2}\mathrm{tr}((\hat{\mathbf{W}}^{(l)} + \hat{\boldsymbol{\mu}}^{(l)}\hat{\boldsymbol{\mu}}^{(l)\mathrm{T}})\mathbf{W}^{-1}) - \frac{1}{2}\ln(\det\boldsymbol{\Sigma}) + \text{ const.} \tag{20}$$

with $\mathrm{tr}(\hat{\boldsymbol{\mu}}^{(l)}\hat{\boldsymbol{\mu}}^{(l)\mathrm{T}}\mathbf{W}^{-1}) = \hat{\boldsymbol{\mu}}^{(l)\mathrm{T}}\mathbf{W}^{-1}\hat{\boldsymbol{\mu}}^{(l)}$ and using the notation $\mathrm{tr}(A)$ for the trace of a matrix A.

Thus, for any fixed $\boldsymbol{\tau}$, we get the expectations needed in the E-step in closed-form expressions. Then we turn to the M-step and maximize $q(\boldsymbol{\varphi}, \hat{\boldsymbol{\varphi}}^{(l)}; \boldsymbol{\tau})$ with respect to the model parameters $\boldsymbol{\varphi}$**.** This expectation-maximization step is achieved for any fix $\boldsymbol{\tau}$. Then, the final Variational EM (VEM) loop is completed by adding an updating step for the variational parameter $\boldsymbol{\tau}$.

To precise the procedure, let us denote now $\mathbf{M}(\boldsymbol{\tau}, \boldsymbol{\beta})$ for $\mathbf{M}$, $\mathbf{W}(\boldsymbol{\tau}, \boldsymbol{\Sigma})$ for $\mathbf{W}$, and $\boldsymbol{\mu}(\boldsymbol{\tau}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \mathbf{W}(\boldsymbol{\tau}, \boldsymbol{\Sigma})\mathbf{M}(\boldsymbol{\tau}, \boldsymbol{\beta})$.

Starting with an initialization $\hat{\boldsymbol{\varphi}}^{(0)}, \hat{\boldsymbol{\tau}}^{(0)}$, the l-th iteration of the algorithm is achieved in three steps. For $l = 1, 2, ...$ we follow the procedure hereafter:

1. E-step. Compute $q(\boldsymbol{\varphi}, \hat{\boldsymbol{\varphi}}^{(l-1)}; \hat{\boldsymbol{\tau}}^{(l-1)})$ defined by (20). In particular, compute $\hat{\mathbf{W}}^{(l-1)} = \mathbf{W}(\hat{\boldsymbol{\tau}}^{(l-1)}, \hat{\boldsymbol{\Sigma}}^{(l-1)})$, $\hat{\mathbf{M}}^{(l-1)} = \mathbf{M}(\hat{\tau}^{(l-1)}, \hat{\boldsymbol{\beta}}^{(l-1)})$ and $\hat{\boldsymbol{\mu}}_1^{(l-1)} = \hat{\mathbf{W}}^{(l-1)}\hat{\mathbf{M}}^{(l-1)}$.
2. M-step for the model parameters.
   (a) Compute $\hat{\boldsymbol{\beta}}^{(l)} = \arg\max_{\boldsymbol{\beta}} \left(T_2(\hat{\boldsymbol{\tau}}^{(l-1)}, \boldsymbol{\beta}) + \hat{\boldsymbol{\mu}}^{(l-1)\,\mathrm{T}} \mathbf{M}(\hat{\boldsymbol{\tau}}^{(l-1)}, \boldsymbol{\beta}) \right)$
   (b) Update $\hat{\boldsymbol{\beta}}^{(l)}$ in the objective function: compute $\hat{\mathbf{M}}(\hat{\boldsymbol{\tau}}^{(l-1)}, \hat{\boldsymbol{\beta}}^{(l)})$ and $\hat{\boldsymbol{\mu}}_2^{(l-1)} = \hat{\mathbf{W}}^{(l-1)}\hat{\mathbf{M}}(\hat{\tau}^{(l-1)}, \hat{\boldsymbol{\beta}}^{(l)})$. Then compute $\hat{\boldsymbol{\Sigma}}^{(l)} = \arg\max_{\boldsymbol{\Sigma}} \left\{ -\frac{1}{2}\mathrm{tr}((\hat{\mathbf{W}}^{(l-1)} + \hat{\boldsymbol{\mu}}_2^{(l-1)}\hat{\boldsymbol{\mu}}_2^{(l-1)\,\mathrm{T}})\mathbf{W}^{-1}(\hat{\boldsymbol{\tau}}^{(l-1)}, \boldsymbol{\Sigma})) - \frac{1}{2}\ln(\det \boldsymbol{\Sigma}) \right\}$.
3. Variational parameter update:
   Update $\hat{\boldsymbol{\Sigma}}^{(l)}$ in $\hat{\mathbf{W}}(\hat{\boldsymbol{\tau}}^{(l-1)}, \hat{\boldsymbol{\Sigma}}^{(l)})$ and $\hat{\boldsymbol{\mu}}_3^{(l-1)} = \hat{\boldsymbol{\mu}}(\hat{\tau}^{(l-1)}, \hat{\boldsymbol{\beta}}^{(l)}, \hat{\boldsymbol{\Sigma}}^{(l)})$,
   Then compute $\hat{\boldsymbol{\tau}}^{(l)} = \arg\max_{\tau} \left\{ \begin{array}{c} T_1(\boldsymbol{\tau}) + T_2(\boldsymbol{\tau}, \hat{\boldsymbol{\beta}}^{(l)}) + \hat{\boldsymbol{\mu}}_3^{(l-1)\,\mathrm{T}} \boldsymbol{M}(\boldsymbol{\tau}, \hat{\boldsymbol{\beta}}^{(l)}) \\ -\frac{1}{2}\mathrm{tr}((\hat{\mathbf{W}}(\hat{\boldsymbol{\tau}}^{(l-1)}, \hat{\boldsymbol{\Sigma}}^{(l)}) + \hat{\boldsymbol{\mu}}_3^{(l-1)}\hat{\boldsymbol{\mu}}_3^{(l-1)\,\mathrm{T}})\mathbf{W}^{-1}(\boldsymbol{\tau}, \hat{\boldsymbol{\Sigma}}^{(l)}) \end{array} \right\}$.

Now, let us consider the initialization and steps 2 and 3 in details.

Initialization

We here discuss how to choose starting values for the VEM algorithm. For the simulation study presented in the next section, we just use the true parameter values that were used for simulation. However, we need to initialize the variational parameter. Let us recall that the variational lower bound of the likelihood equals the likelihood for $\boldsymbol{\tau}$ such that $\tau(\mathbf{s})^2 = Y(\mathbf{s})^2$ for each $\mathbf{s} \in D$, where recall $Y(\mathbf{s}) = \mathbf{X}(\mathbf{s})^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon(\mathbf{s})$. Starting with this initial choice would induce $L_c^{(0)}(\mathbf{Z}, \boldsymbol{\varepsilon} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \widetilde{L_c}^{(0)}(\mathbf{Z}, \boldsymbol{\varepsilon} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\tau})$. Then we choose to initialize the algorithm with $\hat{\boldsymbol{\tau}}^0$ defined by $\hat{\tau}^{(0)}(\mathbf{s}) = (\mathbf{X}(\mathbf{s})^{\mathrm{T}}\hat{\boldsymbol{\beta}}^{(0)} + \eta(\mathbf{s})) \times (2Z(\mathbf{s}) - 1)$, where the variables $\eta(\mathbf{s})$ are independent zero mean Gaussian variables with variance 1. Thus we have $\hat{\tau}^{(0)}(\mathbf{s})^2 = ((\mathbf{X}(\mathbf{s})^{\mathrm{T}}\hat{\boldsymbol{\beta}}^{(0)} + \eta(\mathbf{s}))^2$. Adding the value $\eta(\mathbf{s})$ also ensures that $\hat{\tau}^{(0)}(\mathbf{s})$ is not equal to zero which is required to compute $\lambda(\tau(\mathbf{s}))$.

Step 2-a

For any fixed $\boldsymbol{\tau}$ we want to maximise,

$$T(\boldsymbol{\beta}) = \sum_{\mathbf{s}\in D} \left\{ -\lambda(\tau(\mathbf{s}))(\mathbf{X}(\mathbf{s})^{\mathrm{T}}\boldsymbol{\beta})^2 + (\mathbf{X}(\mathbf{s})^{\mathrm{T}}\boldsymbol{\beta})(Z(\mathbf{s}) - \frac{1}{2} - 2\lambda(\tau(\mathbf{s}))\hat{\mu}(\mathbf{s})) \right\},$$

which is a quadratic function of $\boldsymbol{\beta}$.

Let us note $G(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} T(\boldsymbol{\beta}) = \sum_{\mathbf{s}\in D} \left(-2\lambda(\tau(\mathbf{s}))(\mathbf{X}(\mathbf{s})^{\mathrm{T}}\boldsymbol{\beta}) + Z(\mathbf{s}) - \frac{1}{2} - 2\lambda(\tau(\mathbf{s}))\hat{\mu}(\mathbf{s}) \right) \mathbf{X}(\mathbf{s})$; if the dimension of $\boldsymbol{\beta}$ is 1 or 2, we can solve $G(\boldsymbol{\beta}) = 0$ easily; otherwise we choose to use a Newton-Raphson algorithm, that is, we solve

$$\hat{\boldsymbol{\beta}}^{(k)} = \hat{\boldsymbol{\beta}}^{(k-1)} - \left(\frac{\partial}{\partial \boldsymbol{\beta}} G(\boldsymbol{\beta}) \right)^{-1}_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^{(k-1)}} G(\hat{\boldsymbol{\beta}}^{(k-1)})),$$

with $\frac{\partial}{\partial \boldsymbol{\beta}} G(\boldsymbol{\beta}) = \sum_{\mathbf{s}\in D} -2\lambda(\tau(\mathbf{s}))\mathbf{X}(\mathbf{s})\mathbf{X}(\mathbf{s})^{\mathrm{T}}$, until $\hat{\boldsymbol{\beta}}^{(k)} \simeq \hat{\boldsymbol{\beta}}^{(k-1)}$ and we take $\hat{\boldsymbol{\beta}}^{(l)} = \hat{\boldsymbol{\beta}}^{(k)}$.

Step 2-b

Since $\mathbf{W}^{-1} = \boldsymbol{\Sigma}^{-1} + 2\boldsymbol{\Lambda}$, and for fixed $\boldsymbol{\tau}$ and $\boldsymbol{\beta}$, we search for

$$\hat{\boldsymbol{\Sigma}}^{(l)} = \arg\max_{\boldsymbol{\Sigma}} \left\{ -\frac{1}{2}\mathrm{tr}((\hat{\mathbf{W}} + \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^{\mathrm{T}})\boldsymbol{\Sigma}^{-1}) - \frac{1}{2}\ln(\det \boldsymbol{\Sigma}) \right\}.$$

Writing the covariance matrix as $\mathbf{\Sigma} = \sigma_\varepsilon^2 \mathbf{Q}$, we want to minimise:

$$f(\mathbf{Q}, \sigma_\varepsilon^2) = \frac{1}{\sigma_\varepsilon^2} \mathrm{tr}((\hat{\mathbf{W}} + \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^{\mathrm{T}})\mathbf{Q}^{-1}) + n \ln \sigma_\varepsilon^2 + \ln(\det \mathbf{Q}),$$

with respect to σ_ε^2 and $\mathbf{Q}$. The derivative with respect to σ_ε^2 for a fixed $\mathbf{Q}$ gives the following explicit solution:

$$\sigma_\varepsilon^2(\mathbf{Q}) = \frac{1}{n} \mathrm{tr}((\hat{\mathbf{W}} + \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^{\mathrm{T}})\mathbf{Q}^{-1}). \tag{21}$$

Then the M-step is to minimize, with respect to $\mathbf{Q}$,

$$g(\mathbf{Q}) = n \ln \left[\frac{1}{n} \mathrm{tr}((\hat{\mathbf{W}} + \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^{\mathrm{T}})\mathbf{Q}^{-1}) \right] + \ln(\det \mathbf{Q}).$$

If we assume a parametric feature for $\mathbf{Q}$, then we write $\mathbf{Q} = \mathbf{Q}(\boldsymbol{\theta})$ and the minimization above is with respect to parameter $\boldsymbol{\theta}$. For instance, we can choose the exponential covariance function to characterize the spatial covariance matrix $\mathbf{\Sigma}$; that is, $\mathbf{\Sigma} = (\Sigma_{ij})$ with $\Sigma_{ij} = C(\mathbf{s}_i - \mathbf{s}_j)$ and $C(\mathbf{h}) = \sigma^2 e^{-\|\mathbf{h}\|/\theta}$, for $\mathbf{h} \in \mathbb{R}^2$; we search for a scalar parameter θ in this case.

Step 3

Let us denote $\hat{W}_{ss}$ the s-th diagonal element of $\hat{\mathbf{W}}$, $\hat{K}^{(l)} = (X(\mathbf{s})^{\mathrm{T}}\hat{\boldsymbol{\beta}}^{(l)})^2 + 2\hat{\mu}(\mathbf{s})(\mathbf{X}(\mathbf{s})^{\mathrm{T}}\hat{\boldsymbol{\beta}}^{(l)}) + \hat{W}_{ss} + \hat{\mu}(\mathbf{s})^2$, and $A(\hat{\boldsymbol{\varphi}}^{(l)}; x) = \ln g(x) - \dfrac{x}{2} - \lambda(x)[\hat{K}^{(l)} - x^2]$. Then we write,

$$E\left[\widetilde{L}_c(\mathbf{Z}, \boldsymbol{\varepsilon} \mid \boldsymbol{\varphi}, \boldsymbol{\tau}) \mid \mathbf{Z}, \hat{\boldsymbol{\varphi}}^{(l)}\right] = A(\hat{\boldsymbol{\varphi}}^{(l)}; \boldsymbol{\tau}) \ + \ \text{other terms that do not depend on } \boldsymbol{\tau}.$$

Recall that $\lambda(x) = \dfrac{g(x) - 1/2}{2x}$, and notice that $(\ln g(x))^{\mathrm{T}} = g(-x)$. Then, $\dfrac{\partial}{\partial x} A(\hat{\boldsymbol{\varphi}}^{(l)}; x) = g(-x) - \dfrac{1}{2} - \lambda^{\mathrm{T}}(x)\left[\hat{K}^{(l)} - x^2\right] + 2x\lambda(x)$.

A simple calculus leads to $2x\lambda(x) + g(-x) - \frac{1}{2} = g(x) + g(-x) - 1 = 0$ and $\dfrac{\partial}{\partial x} A(\hat{\boldsymbol{\varphi}}^{(l)}; x) = -\lambda^{\mathrm{T}}(x)\left[\hat{K}^{(l)} - x^2\right]$.

Let us compute $\lambda'(x) = \dfrac{e^{-x}}{4x^2(1 + e^{-x})^2} f(x)$, where $f(x) = 2x - e^x + e^{-x}$. From $f'(x) = -(e^{x/2} - e^{-x/2})^2$, we see that λ' has no zeros; we deduce that $\dfrac{\partial}{\partial x} A(\hat{\boldsymbol{\varphi}}^{(l)}; x) = 0$ for $x^2 = \hat{K}^{(l)}$.

Then we get a closed-form expression to update the variational parameter, that is,

$$\hat{\tau}^{(l)}(\mathbf{s})^2 = (\mathbf{X}(\mathbf{s})^{\mathrm{T}}\hat{\boldsymbol{\beta}}^{(l)})^2 + 2(\mathbf{X}(\mathbf{s})^{\mathrm{T}}\hat{\boldsymbol{\beta}}^{(l)})\hat{\mu}^{(l)}(\mathbf{s}) + \hat{W}_{ss}^{(l)} + \hat{\mu}^{(l)}(\mathbf{s})^2. \tag{22}$$

This result is not surprising; recall the inequality (9); we have equality between both sides if $x^2 = \tau^2$; in other words, we are looking for $\tau(\mathbf{s})^2$ as close as possible to $Y(\mathbf{s})^2 = ((\mathbf{X}(\mathbf{s})^{\mathrm{T}}\boldsymbol{\beta}) + \varepsilon(\mathbf{s}))^2$; then we take $\hat{\tau}^{(l)}(\mathbf{s})^2 = E[Y(\mathbf{s})^2 \mid \mathbf{Z}, \hat{\boldsymbol{\varphi}}^{(l-1)}]$ which is exactly our result.

Finally, we update for each $\mathbf{s} \in D$,

$$\hat{\tau}^{(l)}(\mathbf{s}) = \sqrt{\hat{\tau}^{(l)}(\mathbf{s})^2} \times (2Z(\mathbf{s}) - 1).$$

4. Experiments

In this Section, we conduct simulations and run the VEM algorithm described in the previous Section to derive model parameter estimates. We compare our method with Laplace approximations.

Let D be a square lattice of size 40×60, with $n = 2400$ sites. Following the model's description in Section 2, we start by simulating a Gaussian random field with spatial covariance matrix $\boldsymbol{\Sigma}$**.** Then we simulate independent Bernoulli random variables.

The Gaussian random field $\boldsymbol{\varepsilon}$ is simulated on D with distribution $N_n(\mathbf{0}; \boldsymbol{\Sigma})$; we choose the exponential covariance function to characterize the spatial covariance matrix; that is, $\boldsymbol{\Sigma} = (\Sigma_{ij})_{i,j}$ with $\Sigma_{ij} = C(\mathbf{s}_i - \mathbf{s}_j)$ and $C(\mathbf{h}) = \sigma^2 e^{-\frac{\|\mathbf{h}\|}{\theta}}$, for $\mathbf{h} \in \mathbb{R}^2$. In order to obtain reasonable spatial dependence, we choose $\theta = 5$ and then $\theta = 15$, the latter value ensuring stronger spatial dependence. We set $\sigma^2 = 1$.

We choose the trend to be linked to the spatial location on D; for $\mathbf{s} = (s_1, s_2) \in D$,

$$X(\mathbf{s})^{\mathrm{T}}\boldsymbol{\beta} = (1, s_1 - 20, s_2 - 30)(\beta_0, \beta_1, \beta_2)^{\mathrm{T}}.$$

Now, let us define the variation of the 'signal', V_s, as $V_s = \frac{1}{n}\mathrm{tr}(\boldsymbol{\Sigma}) + \frac{1}{n}\sum_{i=1}^{n}\left(X(\mathbf{s}_i)^{\mathrm{T}}\boldsymbol{\beta} - \mathrm{average}_{\mathbf{s}\in D}(X(\mathbf{s})^{\mathrm{T}}\boldsymbol{\beta})\right)^2$. Following Aldworth and Cressie (1999), the parameter $\boldsymbol{\beta}$ is selected such that V_s is approximately 2. Here we specify $\beta_0 = \frac{1}{10}$, $\beta_1 = \frac{1}{16}$ and $\beta_2 = \frac{1}{24}$ which gives $V_s \simeq 2$, and balances the effect of β_1 and β_2 (β_0 is a free parameter that does not impact V_s).

We next compute $Y(\mathbf{s}_1), \cdots, Y(\mathbf{s}_n)$ defined in (3); then, conditionally to $\mathbf{Y}$, we simulate independent Bernoulli random variables $Z(\mathbf{s})$, with success probabilities defined in (2), $p(\mathbf{s}) = \dfrac{e^{Y(\mathbf{s})}}{1 + e^{Y(\mathbf{s})}}$.

Each model is simulated $L = 100$ times, as described above. Then estimation is performed on each simulation based on the procedure described in Section 3. We also compute the estimates of the parameters obtained from the Laplace approximation procedure; considering the complete likelihood, since the expectations in the E-step of the EM algorithm are not available in closed form, we use Laplace approximations to approximate the intractable integrals. Laplace approximations are based on second-order Taylor-series expansions of the integrands around the mode, see for instance Sengupta and Cressie (2013); we give hereafter the main result, details of the calculation can be found in the Appendix. The issue is to calculate, at the $(k+1)$-th iteration of the EM algorithm, $q(\boldsymbol{\varphi}, \hat{\boldsymbol{\varphi}}^{(k)}) = E[L_c(\mathbf{Z}, \boldsymbol{\varepsilon} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}) \mid \mathbf{Z}, \hat{\boldsymbol{\beta}}^{(k)}, \hat{\boldsymbol{\Sigma}}^{(k)}]$. The Laplace approximation of this expectation is $\tilde{q}(\boldsymbol{\varphi}, \hat{\boldsymbol{\varphi}}^{(k)})$ defined as,

$$\begin{aligned}\tilde{q}(\boldsymbol{\varphi}, \hat{\boldsymbol{\varphi}}^{(k)}) = \sum_{\mathbf{s}\in D}\left[-\ln(1 + e^{Y_m^{(k)}(\mathbf{s})}) + \frac{1}{2}\frac{e^{Y_m^{(k)}(\mathbf{s})}}{(1 + e^{Y_m^{(k)}(\mathbf{s})})^2}(H(\boldsymbol{\varepsilon}_m^{(k)})^{-1})_{ss} + Y_m^{(k)}(\mathbf{s})Z(\mathbf{s})\right] \\ - \frac{1}{2}\ln(\det\Sigma) - \frac{1}{2}\boldsymbol{\varepsilon}_m^{(k)\,\mathrm{T}}\Sigma^{-1}\boldsymbol{\varepsilon}_m^{(k)} - \frac{1}{2}\mathrm{tr}(\Sigma^{-1}(-H(\boldsymbol{\varepsilon}_m^{(k)})^{-1}) - \frac{n}{2}\ln(2\pi). \qquad (23)\end{aligned}$$

where $\boldsymbol{\varepsilon}_m^{(k)}$ is the mode of $L_c(\mathbf{Z}, \boldsymbol{\varepsilon} \mid \hat{\boldsymbol{\beta}}^{(k)}, \hat{\boldsymbol{\Sigma}}^{(k)})$, $H(\boldsymbol{\varepsilon}_m^{(k)})$ is the Hessian computed at the mode, and $Y_m^{(k)}(\mathbf{s}) = \mathbf{X}(\mathbf{s})^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_m^{(k)}(\mathbf{s})$.

We display in Table 1 and Table 2 the means and mean square errors (MSE) of the estimates, for both methods, obtained from the 100 simulations.

	β_0	β_1	β_2	σ^2	θ
Target	0.1	0.0625	0.0417	1	5
Variational method	0.0610	0.0596	0.0406	0.6378	5.2309
MSE	0.0322	0.0002	0.0001	0.1548	2.1402
Laplace approximation	0.0828	0.0477	0.0320	0.8751	4.5952
MSE	0.0013	0.0004	0.0001	0.0429	1.4480

Table 1: Mean and MSE of VEM and Laplace estimates, $\theta = 5$

	β_0	β_1	β_2	σ^2	θ
Target	0.1	0.0625	0.0417	1	15
Variational method	0.1162	0.0607	0.0408	0.8146	12.8138
MSE	0.0026	$4.\ 10^{-5}$	$2.\ 10^{-5}$	0.1499	35.0405
Laplace approximation	0.0837	0.0541	0.0358	0.9984	15.1183
MSE	0.0068	0.0001	0.0001	0.3119	79.4734

Table 2: Mean and MSE of VEM and Laplace estimates, $\theta = 15$

In the case of weak spatial dependence, $\theta = 5$, we observe a negative bias for $\boldsymbol{\beta}$ and σ^2 for both methods; then we observe a positive bias for the VEM estimate of θ, and a negative bias for the Laplace estimate. The MSE computed for VEM estimates are quite good, but they are greater or equal than those computed for Laplace estimates. However, the difference between the MSE of both methods is not very large.

On the other hand, in case of stronger spatial dependence, the VEM method performs better. For $\theta = 15$, we observe a negative bias for $\boldsymbol{\beta}$ and σ^2 for Laplace estimates, and a positive bias for the estimate of θ. While the bias is either positive or negative for the VEM estimates. More important, the MSE computed for VEM estimates are quite good; we notice that they are less than 0.0001 for parameters β_1 and β_2. For the other parameters they are about half the MSE computed for the Laplace estimates. Especially for parameter θ, the MSE of the Laplace estimate is quite large, because the method sometimes completely fails and proposes an absurd value for this parameter.

As stated before, there are no theoretical results on the variance of the variational estimator. Hence we follow a parametric bootstrap approach as described hereafter to approximate the variance. Let us note $(\boldsymbol{\beta}^\star, \sigma^{2\star}, \theta^\star)$ a set of estimates resulting from the VEM procedure. We simulate $\boldsymbol{\varepsilon}^{\star(b)}$ and $\mathbf{Z}^{\star(b)}$, B times, for $b = 1, \ldots, B$, using $(\boldsymbol{\beta}^\star, \sigma^{2\star}, \theta^\star)$ as simulation parameters. For each new simulated data set $\mathbf{Z}^{\star(b)}$, we compute the VEM estimates $(\hat{\boldsymbol{\beta}}^{\star(b)}, \hat{\sigma}^{2\star(b)}, \hat{\theta}^{\star(b)})$; then, the bootstrap variance of the VEM estimator is given by the empirical variance of the B estimates $(\hat{\boldsymbol{\beta}}^{\star(b)}, \hat{\sigma}^{2\star(b)}, \hat{\theta}^{\star(b)})$, see Beran (2003). We choose $B = 150$ and we consider two realisations from the previous results; we consider the set $(\boldsymbol{\beta}^\star, \sigma^{2\star}, \theta^\star) = (0.1042, 0.0648, 0.0429, 0.9537, 14.9845)$, for which $\sigma^{2\star}$ and $\theta^\star$ are very close to $\sigma^2 = 1$ and $\theta = 15$; then we consider the second set $(\boldsymbol{\beta}^\star, \sigma^{2\star}, \theta^\star) = (0.1387, 0.0649, 0.0375, 0.8246, 12.818)$, for which $\sigma^{2\star}$ and $\theta^\star$ are close to the mean values of the estimates of $\sigma^2 = 1$ and $\theta = 15$ given in Table 2. We make this choice because our main interest is on the spatial dependence parameters. We do the same work with the Laplace estimates, and we present the results in Table 3. We obtain very similar

	$\beta_0^\star$	$\beta_1^\star$	$\beta_2^\star$	$\sigma^{2\star}$	$\theta^\star$
VEM estimate	0.1042	0.0648	0.0429	0.9537	14.9845
Bootstrap std	0.0233	0.0031	0.0020	0.2303	3.8165
Laplace estimate	0.1162	0.0545	0.0390	1.0011	15.0463
Bootstrap std	0.0273	0.0020	0.0010	0.2798	4.2371

Table 3: Bootstrap standard deviation of the VEM and Laplace estimators

bootstrap variance values for the two trials $(\boldsymbol{\beta}^\star, \sigma^{2\star}, \theta^\star)$, in each case, VEM or Laplace; hence we display the results for only one set. We note that the bootstrap standard deviations are slightly smaller for the trend estimates $\hat{\boldsymbol{\beta}}^\star$ resulting from the Laplace procedure than those of the VEM estimates; on the contrary, looking at the spatial dependence parameters, the bootstrap standard deviations of the VEM estimates are smaller than the ones of the Laplace estimates.

We notice that the bias on the covariance parameters is a bit large; in order to investigate the effect of the lattice size on the bias, we run other simulations with $\sigma^2 = 1$ and $\theta = 5$ considering lattice sizes $n = 30 \times 30$, $n = 40 \times 40$, then $n = 60 \times 60$. In each case, we adapt $X(\mathbf{s}) = (1, s_1 - \frac{\sqrt{n}}{2}, s_2 - \frac{\sqrt{n}}{2})$ and the parameter $\boldsymbol{\beta} = (\frac{1}{10}, \beta_1, \beta_2)$ in order to keep $V_s \simeq 2$; we present the results in Table 4. Obviously, the standard deviation tends to decrease when n increases; in most cases the bias is also reduced. The bias of parameter θ is larger for $n = 3600$ than for $n = 2400$ (but the standard deviation is reduced); an explanation is that a scale value of 5 characterises a weak spatial dependence in this case, weaker for the larger lattice; it might be hard to detect it correctly on some simulations. Let us note that the number of iterations and processing time of the algorithm both increase with n; we observe the same phenomenon for the EM algorithm with Laplace approximations; the algorithms are slowed down by the size of the involved matrices, but also seem to have difficulty to reach the optimum value of the log likelihood, the log likelihood value evolving slightly. Thus, for large lattices, we do recommend to use an approach which avoids the computation of large dimension matrices, as discussed in Section 6.

Let us compare the variational and Laplace methods with respect to the processing time; the Laplace approximation method particularly requires to compute the mode $\boldsymbol{\varepsilon}_m$ maximising $L_c(\boldsymbol{Z}, \boldsymbol{\varepsilon} \mid \boldsymbol{\varphi})$ at each iteration of the EM algorithm, and the Hessian matrix; the Variational method ignores this stage but adds the updating step of the variational parameters. However, the Variational EM remains faster than the Laplace EM. Especially for large θ, the time processing for computing the Laplace estimates becomes important, while it does not increase for the VEM method. For example, the average time for one iteration of the VEM algorithm for $\theta = 5$ is 39.08 seconds, and 45.22 seconds for the Laplace EM; for $\theta = 15$, the difference is a bit larger, with 47.77 seconds for the VEM algorithm and 54.74 for the Laplace EM. The average number of iterations is similar for both methods, 3.16 for the VEM algorithm and 3.00 for the Laplace EM, for $\theta = 5$. Finally, we notice that for the Laplace EM, we sometimes get weird results, completely out or range estimates for σ^2 and θ, while the VEM leads to more regular values.

Finally, we investigate how sensitive are the estimates to the initial value of the algorithm. This study has been conducted for a lattice size of 40×60 and true parameter values $\boldsymbol{\beta} = (0.1, 0.0625, 0.0417)$, $\sigma^2 = 1$ and $\theta = 15$. Here, we generate random initial values of the

	β_0	β_1	β_2	σ^2	θ
Target $n = 900$	0.1	0.0833	0.0833	1	5
Mean	0.1292	0.0822	0.0861	0.6278	4.1865
Std	0.2279	0.0238	0.0198	0.2326	1.6134
Target $n = 1600$	0.1	0.0625	0.0625	1	5
Mean	0.1034	0.0607	0.0580	0.6284	4.7605
Std	0.2024	0.0142	0.0146	0.1709	1.3341
Target $n = 2400$	0.1	0.0625	0.0417	1	5
Mean	0.0610	0.0596	0.0406	0.6378	5.2309
Std	0.1752	0.0125	0.0116	0.1538	1.4446
Target $n = 3600$	0.1	0.0417	0.0417	1	5
Mean	0.1149	0.0397	0.0415	0.6720	5.5579
Std	0.1690	0.0080	0.0079	0.1287	1.1918

Table 4: Mean and Standard deviation of VEM estimates for increasing lattice sizes, $\sigma^2 = 1$ and $\theta = 5$

model parameters $(\hat{\boldsymbol{\beta}}^{(0)}, \hat{\sigma}^{2(0)}, \hat{\theta}^{(0)})$ in the estimation procedure; we recall that the initial value of the variational parameter $\hat{\boldsymbol{\tau}}^{(0)}$ is given by $\hat{\tau}^{(0)}(\mathbf{s}) = (\mathbf{X}(\mathbf{s})^{\mathrm{T}}\hat{\boldsymbol{\beta}}^{(0)} + \eta(\mathbf{s})) \times (2Z(\mathbf{s}) - 1)$, where the $\eta(\mathbf{s})$ are i.i.d. $N(0, 1)$. We use the same random initial values for both Variational and Laplace procedures. The first remark is that in most cases, the Laplace algorithm fails and stops, usually at the step of finding the mode $\boldsymbol{\varepsilon}_m$, while the VEM algorithm always gives a final result. Of course, the number of iterations of the VEM algorithm is quite large. We observe that the final estimate values for the trend parameter $\boldsymbol{\beta}$ are usually not so far from the initial values. The mean values are $\bar{\hat{\boldsymbol{\beta}}} = (2.0150, 1.2251, 1.4337)$ with standard deviations $(2.49, 1.04, 1.16)$. At least, we do not observe large outliers leading to estimates ten times larger or smaller than the target values. But the spatial dependence estimates $\hat{\sigma}^2$ and $\hat{\theta}$ are more sensitive to the starting values. Then we run other experiments, starting with the true parameter values for $\hat{\boldsymbol{\beta}}^{(0)}$ and different values for $\hat{\sigma}^{2(0)}$ and $\hat{\theta}^{(0)}$. We observe that for small data sets, the algorithm converge to the correct values. For larger data sets, we obtain close estimate values for $\hat{\sigma}^2$, but the final estimates of $\hat{\theta}$ often remain close to the starting value. For instance, starting from $\hat{\sigma}^{2(0)} = 2$ and $\hat{\theta}^{(0)} = 5$, we obtain mean values $\hat{\sigma}^2 = 0.6044$ and $\hat{\theta} = 4.8307$; but the likelihood values are much less than the one obtained starting with the true values. To conclude this experimental study, let us note that for real datasets, we propose a method for choosing initial values, that we present in the next section.

5. Application to a real data set

We consider the study of a real data set; the columbus data is available in the R package spdep. The data concerns 49 neighbourhoods in Columbus, Ohio, United States. Together with location variables, the data also records the following variables: CRIME, residential burglaries and vehicle thefts per thousand households in the neighbourhood, HOVAL, housing value (in \$1,000), and INC, the household income (in \$1,000). From the variable CRIME, we form the binary variable CRIME2, which takes the value 1 if the value CRIME is over the median value, that is 34, and 0 otherwise. We consider HOVAL, INC, and X and Y, the coordinates of the neighbourhood centres, as covariates.

	$\beta_0^\star$	$\beta_1^\star$	$\sigma^{2\star}$	$\theta^\star$
VEM estimate	5.8652	−0.4218	0.0493	2.5353
Bootstrap std	1.5400	0.1148	0.0045	0.3323
GLM Standard deviation	1.6127	0.1163		

Table 5: Bootstrap standard deviation of the VEM estimates and GLM standard deviations

When dealing with simulated data, we take for starting values of the model parameters in the EM algorithm the true values that were used for simulation. We have discussed in Section 3 the initialization of the variational parameter, which is also related to the starting values of the model parameter $\hat{\boldsymbol{\beta}}^{(0)}$. For real data applications, we propose the following procedure. We run an ordinary GLM model (with no random effects) for CRIME2, with our covariates as explanatory variables; we run all possible embedded models and select the best one according to AIC and BIC criteria. In our case, the model with the lowest AIC and BIC values was obtained with the single covariate INC. Thus we consider the following model (3):

$$Y(\mathbf{s}) = \mathbf{X}(\mathbf{s})^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon(\mathbf{s})$$

with $\mathbf{X}(\mathbf{s})^{\mathrm{T}}\boldsymbol{\beta} = (1, INC(\mathbf{s}))(\beta_0, \beta_1)^{\mathrm{T}}$.

The starting values for parameter $\boldsymbol{\beta}$ are obtained by the ordinary GLM procedure, $\hat{\boldsymbol{\beta}}^{(0)} = \hat{\boldsymbol{\beta}}_{GLM} = (5.8877994, -0.4226277)^{\mathrm{T}}$. This also allows to compute the starting values $\hat{\tau}^{(0)}(\mathbf{s}) = (\mathbf{X}(\mathbf{s})^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{GLM} + \eta(\mathbf{s})) \times (2Z(\mathbf{s}) - 1)$.

Furthermore, we need starting values for the covariance parameters, as well as the parametric feature of the spatial covariance $\boldsymbol{\Sigma}$ of $\boldsymbol{\varepsilon}$. We write $\varepsilon(\mathbf{s}) = Y(\mathbf{s}) - \mathbf{X}(\mathbf{s})^{\mathrm{T}}\boldsymbol{\beta}$, and recall that $Y(\mathbf{s}) = \log \dfrac{p(\mathbf{s})}{1 - p(\mathbf{s})}$. Then, define $U(\mathbf{s}) = \log \dfrac{\bar{\mathbf{Z}}}{1 - \bar{\mathbf{Z}}} - \mathbf{X}(\mathbf{s})^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{GLM}$, with $\bar{\mathbf{Z}} = \frac{1}{n}\sum_{\mathbf{s}\in D} Z(\mathbf{s})$. We compute the variogram of $\mathbf{U}$ and fit the latter, with different models. Here, the best fit was obtained for the exponential model, without nugget effect, and parameters $\sigma^2 = 7.608678$ and $\theta = 6.152822$. Hence, we choose the exponential model for the covariance matrix $\boldsymbol{\Sigma}$, and we use the previous values as starting values in the VEM algorithm. We finally obtain the following estimates, $\hat{\boldsymbol{\beta}}_{VEM} = (5.8652, -0.4218)^{\mathrm{T}}$, $\hat{\sigma}^2_{VEM} = 0.0493$ and $\hat{\theta}_{VEM} = 2.5353$. We note that if the final estimate $\hat{\boldsymbol{\beta}}_{VEM}$ is close to the initial $\hat{\boldsymbol{\beta}}_{GLM}$, this is not at all the case for $\hat{\sigma}^2_{VEM}$ and $\hat{\theta}_{VEM}$. In order to check the sensitivity to the initialization, we run again the algorithm for other starting values $\hat{\sigma}^{2(0)}$ and $\hat{\theta}^{(0)}$, for instance $\hat{\sigma}^{2(0)} = 1$ and $\hat{\theta}^{(0)} = 10$, and satisfactory enough, we obtained the same result.

We end the study by computing the variance of our estimators by a parametric bootstrap approach, as described in the previous Section. The bootstrap standard deviations are given in Table 5. As a comparison for the trend parameters, we also present the standard deviation of the GLM estimates of the ordinary logistic regression. The bootstrap standard deviations of the VEM estimates are slightly lower than the GLM standard deviations.

6. Discussion and conclusions

In this paper, we have developed a variational parameter estimation procedure for logistic spatial regression. In a classical hierarchical framework, the binary process is obtained from a

hidden Gaussian spatial process together with covariates, via the logit function link. We present in detail the variational estimation method for this model and show its advantages; it bypasses the problematic term $\sum_{\mathbf{s}\in D} \ln(1 + e^{Y(\mathbf{s})})$ issued from the logit function. The variational transformation leads to a lower bound of the log likelihood, that has a Gaussian form. Accordingly, the expectations needed in the E-step are available in closed-form expressions, and do not require a Monte Carlo procedure. The VEM algorithm is easy to implement, it allows fast estimation, and compared to the Laplace approximations, avoids the computation of the mode at each iteration. It is less sensitive to the initialization of the parameters. We have shown through simulations that the VEM method performs better than Laplace approximations in the case of strong spatial dependence. We computed an approximation of the variance of both Laplace and VEM estimators via a bootstrap approach; again, the VEM estimators performs better from this point of view. We also investigated the behaviour of the estimates with respect to the size of the data.

Finally, we conducted a study on a real data set and explained the full procedure to initialize the algorithm, and obtain estimates.

The estimation procedure requires to compute the inverse covariance matrix $\boldsymbol{\Sigma}^{-1}$, which becomes problematic for large data sets. There are several ways to overcome this issue; one can model directly the inverse covariance matrix (see for instance Lindgren et al. (2011)); or we can use a reduced-rank approach (e.g. Wikle (2010)); particularly, we can model the spatial process $\boldsymbol{\varepsilon}$ with a Spatial Random Effects (SRE) model, as described by Cressie and Johanesson (2008), see also Sengupta and Cressie (2013). Then the VEM algorithm has to be adapted to the new writing of the likelihood. This extension is a work in progress.

Appendix

We now derive Laplace approximations to approximate the E-step in (8), which are based on second-order Taylor series expansions of the logarithm of the integrands around their respective modes. Let us recall the expression of the complete log likelihood given in (7):

$$L_c(\boldsymbol{Z}, \boldsymbol{\varepsilon} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}) = -\sum_{\mathbf{s}\in D} \ln(1 + e^{Y(\mathbf{s})}) + \sum_{\mathbf{s}\in D} Y(\mathbf{s})Z(\mathbf{s}) - \frac{1}{2}\ln(\det \boldsymbol{\Sigma}) - \frac{1}{2}\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\,\boldsymbol{\varepsilon} - \frac{n}{2}\ln 2\pi$$

Let us denote $\boldsymbol{\varepsilon}_m$ as the mode of $L_c(\boldsymbol{Z}, \boldsymbol{\varepsilon} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma})$; then, we write the second-order Taylor series expansion for $L_c(\boldsymbol{Z}, \boldsymbol{\varepsilon} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma})$ around $\boldsymbol{\varepsilon}_m$,

$$\begin{aligned} L_c(\boldsymbol{Z}, \boldsymbol{\varepsilon} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}) = L_c(\boldsymbol{Z}, \boldsymbol{\varepsilon}_m \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}) + (\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_m)^{\mathrm{T}} \frac{\partial}{\partial \boldsymbol{\varepsilon}} L_c(\boldsymbol{Z}, \boldsymbol{\varepsilon} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma})\Big|_{\boldsymbol{\varepsilon}=\boldsymbol{\varepsilon}_m} \\ + \frac{1}{2}(\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_m)^{\mathrm{T}} \frac{\partial^2}{\partial \boldsymbol{\varepsilon}\partial \boldsymbol{\varepsilon}^{\mathrm{T}}} L_c(\boldsymbol{Z}, \boldsymbol{\varepsilon} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma})\Big|_{\boldsymbol{\varepsilon}=\boldsymbol{\varepsilon}_m} (\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_m) + \ \ldots \end{aligned}$$

The second term at the right-hand side is in fact zero, so we get the following writing:

$$L_c(\boldsymbol{Z}, \boldsymbol{\varepsilon} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}) \simeq L_c(\boldsymbol{Z}, \boldsymbol{\varepsilon}_m \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}) - \frac{1}{2}(\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_m)^{\mathrm{T}}(-H(\boldsymbol{\varepsilon}_m))(\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_m),$$

where $H(\boldsymbol{\varepsilon}_m) = \dfrac{\partial^2}{\partial \boldsymbol{\varepsilon}\partial \boldsymbol{\varepsilon}^{\mathrm{T}}} L_c(\boldsymbol{Z}, \boldsymbol{\varepsilon} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma})\Big|_{\boldsymbol{\varepsilon}=\boldsymbol{\varepsilon}_m}$.

We deduce that the probability density function of $[\boldsymbol{\varepsilon} \mid \boldsymbol{Z}, \varphi_\varepsilon]$ is approximately proportional to

$\exp L_c(\mathbf{Z}, \boldsymbol{\varepsilon}_m \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}) \times \exp\left[-\frac{1}{2}(\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_m)^{\mathrm{T}}(-H(\boldsymbol{\varepsilon}_m))(\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_m)\right]$; that is, it is approximately proportional to a Gaussian density. Computing the normalisation constant, we conclude that $E\left[\boldsymbol{\varepsilon} \mid \mathbf{Z}, \varphi_\varepsilon\right] \simeq \boldsymbol{\varepsilon}_m$ and $\mathrm{var}(\boldsymbol{\varepsilon} \mid \mathbf{Z}, \varphi_\varepsilon) \simeq -H(\boldsymbol{\varepsilon}_m)^{-1}$.

It remains to compute the expectation of the term $E[\ln(1 + e^{Y(\mathbf{s})}) \mid \mathbf{Z}, \varphi_\varepsilon]$ in (8); we apply the same method and derive a second-order Taylor-series expansion of $\ln(1 + e^{Y(\mathbf{s})})$ around $\varepsilon_m(\mathbf{s})$; denoting $Y_m(s) = \mathbf{X}(\mathbf{s})^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_m(\mathbf{s})$, we obtain,

$$\ln(1 + e^{Y(\mathbf{s})}) = \ln(1 + e^{Y_m(\mathbf{s})}) + (\varepsilon(\mathbf{s}) - \varepsilon_m(\mathbf{s}))\frac{e^{Y_m(\mathbf{s})}}{1 + e^{Y_m(\mathbf{s})}} + \frac{1}{2}(\varepsilon(\mathbf{s}) - \varepsilon_m(\mathbf{s}))^2\frac{e^{Y_m(\mathbf{s})}}{(1 + e^{Y_m(\mathbf{s})})^2} + \ ...$$

Then we can write the desired expectation as follows,

$$E[\ln(1 + e^{Y(\mathbf{s})}) \mid \mathbf{Z}, \varphi_\varepsilon] \simeq \ln(1 + e^{Y_m(\mathbf{s})}) - \frac{1}{2}\frac{e^{Y_m(\mathbf{s})}}{(1 + e^{Y_m(\mathbf{s})})^2}(H(\boldsymbol{\varepsilon}_m)^{-1})_{ss}.$$

Finally, we obtain the following approximation for the expectation needed in the E-step of the EM algorithm,

$$\begin{aligned} q(\boldsymbol{\varphi}, \hat{\boldsymbol{\varphi}}^{(l)}) &= E\left[L_c(\mathbf{Z}, \boldsymbol{\varepsilon} \mid \boldsymbol{\varphi}) \mid \mathbf{Z}, \hat{\boldsymbol{\varphi}}^{(l)}\right] \\ &\simeq -\sum_{\mathbf{s}\in D}\left(\ln(1 + e^{Y_m(\mathbf{s})}) - \frac{1}{2}\frac{e^{Y_m(\mathbf{s})}}{(1 + e^{Y_m(\mathbf{s})})^2}(H(\boldsymbol{\varepsilon}_m)^{-1})_{ss}\right) + \sum_{\mathbf{s}\in D} Y_m(\mathbf{s})Z(\mathbf{s}) \\ &\quad - \frac{1}{2}\ln(\det\boldsymbol{\Sigma}) - \frac{1}{2}\left(\mathrm{tr}(-\boldsymbol{\Sigma}^{-1}H(\boldsymbol{\varepsilon}_m)^{-1}) + \boldsymbol{\varepsilon}_m^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}_m\right) - \frac{n}{2}\ln 2\pi. \end{aligned}$$

The mode $\boldsymbol{\varepsilon}_m$ and the matrix $H(\boldsymbol{\varepsilon}_m)$ are obtained by a standard procedure. The gradient of $L_c(\mathbf{Z}, \boldsymbol{\varepsilon} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma})$ is given by $\frac{\partial}{\partial\boldsymbol{\varepsilon}}L_c(\mathbf{Z}, \boldsymbol{\varepsilon} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \mathbf{Z} - \mathrm{vec}(\frac{e^{\mathbf{Y}_m}}{1 + e^{\mathbf{Y}_m}}) - \boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}$, and we solve the equation $\frac{\partial}{\partial\boldsymbol{\varepsilon}}L_c(\mathbf{Z}, \boldsymbol{\varepsilon} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}) = 0$ by using a Newton-Raphson algorithm. Then a simple calculation gives the Hessian $H(\boldsymbol{\varepsilon}_m) = -\mathrm{diag}\left(\frac{e^{Y_m}}{(1 + e^{Y_m})^2}\right) - \boldsymbol{\Sigma}^{-1}$.

Acknowledgments

This research was conducted as part of the project Labex MME-DII (ANR11-LBX-0023-01).

References

Aldworth, J., Cressie, N., 1999. Sampling designs and prediction methods for Gaussian spatial processes, in: Ghosh, S. (Ed.), Multivariate Analysis, Designs of Experiments, and Survey Sampling, Markel Dekker, Inc., New York, NY, 1–54.

Beran, R., 2003. The Impact of the Bootstrap on Statistical Algorithms and Theory. Statistical Science 18, n2, 175184.

Cappé, O., Moulines, E., Rydén, T., 2005. Inference in Hidden Markov Models, Springer Series in Statistics, Springer-Verlag New York.

Chib S., Greenberg E., 1995. Understanding the Metropolis algorithm. The American statistician 49 n4, 327-335.

Cressie, N., Johannesson, G., 2008. Fixed Rank Kriging for very large spatial data sets, Journal of the Royal Statistical Society, Series B 70, 209-226.

Dempster, A.P., Laird, N., Rubin, D., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B 39, 1-38.

Diggle, P., Giorgi, E., 2015. Model-Based Geostatistics for Prevalence Mapping in Low-Resource Settings. Journal of the American Statistical Association. 111:515, 1096-1120.

Jaakola T., Jordan M., 2000, Bayesian parameter estimation via variational methods. Statistics and Computing 10, 25-37.

Lindgren, F., Rue, H., Lindstrom, J., 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach, Journal of the Royal Statistical Society, Series B 73, 423-498.

McLachlan, G.J., Krishnan, T., 2008. The EM Algorithm and Extensions, second ed. Wiley-Interscience, New York, NY.

Nong, Y., Du, Q., 2011. Urban Growth Pattern Modeling Using Logistic Regression. Geo-spatial Information Science 14, 62-67.

Paciorek, C.J., 2007. Computational Techniques for Spatial Logistic Regression with Large Datasets. Comput Stat Data Anal. 51(8), 36313653.

Robert, C.P., Casella, G., 2004. Monte Carlo Statistical Methods. Springer, New York, NY.

Roberts, G.O., Rosenthal, J.S., 2001. Optimal Scaling for Various Metropolis-Hastings Algorithms Statistical Science, Vol. 16, No. 4, 351-367.

Peyrard, N., Cros, M.-J., de Givry, S., Franc, A., Robin, S., Sabbadin, R., Schiex, T., Vignes, M., 2018. Exact or approximate inference in graphical models: why the choice is dictated by the treewidth, and how variable elimination can be exploited. https://arxiv.org/pdf/1506.08544.pdf

Rustagi, J., 1976. Variational methods in statistics. New York, Academic press.

Schneider, L., Pontius, R.G., 2001. Modeling land use change in the Ipswich watershed, Massachusetts, USA. Agric Ecosyst Environ 85, 8394.

Sengupta, A., Cressie, N., 2013. Hierarchical statistical modelling of big spatial datasets using the exponential family of distributions, Spatial Statistics 4, 14-44.

Sengupta A., Cressie N., Kahn, B.H., Frey, R., 2016. Predictive Inference for Big, Spatial, NonGaussian Data: MODIS Cloud Data and its ChangeofSupport. Australian and New Zealand Journal of Statistics, 58(1), 15-45.

Serneels, S., Lambin, E.F., 2001. Proximate causes of land-use change in Narok District, Kenya: a spatial statistical model. Agric Ecosyst Environ 85, 6581.

Spiegelhalter, D., Lauritzen, S., 1990. Sequential updating of conditional probabilities on directed graphical structures. Networks 20, 579-605.

Tayyebi A., Delavar M.R., Yazdanpanah M.J., Pijanowski B.C., Saeedi S., Tayyebi A.H. (2010) A Spatial Logistic Regression Model for Simulating Land Use Patterns: A Case Study of the Shiraz Metropolitan Area of Iran, in: Chuvieco E., Li J., Yang X. (Eds.), Advances in Earth Observation of Global Change. Springer, Dordrecht.

Wikle, C.K., 2010. Low rank representations for spatial processes, in: Gelfand, A., Diggle, P., Fuentes, M., Guttorp, P. (Eds.), Handbook of Spatial Statistics. Chapman and Hall, CRC Press, Boca Raton, FL, pp. 107-118.

Wu, W., Zhang, L., 2013. Comparison of spatial and non-spatial logistic regression models for modeling the occurrence of cloud cover in north-eastern Puerto Rico. Applied Geography 37, 52-62.