Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2010)
72, *Part* 3, *pp.* 389–402

On selection of spatial linear models for lattice data

Jun Zhu,

Colorado State University, Fort Collins, and University of Wisconsin, Madison, USA

Hsin-Cheng Huang

Academia Sinica, Taipei, and National Chiao Tung University, Hsinchu, Taiwan

and Perla E. Reyes

University of Wisconsin, Madison, USA

[Received February 2009. Final revision December 2009]

Summary. Spatial linear models are popular for the analysis of data on a spatial lattice, but statistical techniques for selection of covariates and a neighbourhood structure are limited. Here we develop new methodology for simultaneous model selection and parameter estimation via penalized maximum likelihood under a spatial adaptive lasso. A computationally efficient algorithm is devised for obtaining approximate penalized maximum likelihood estimates. Asymptotic properties of penalized maximum likelihood estimates and their approximations are established. A simulation study shows that the method proposed has sound finite sample properties and, for illustration, we analyse an ecological data set in western Canada.

Keywords: Conditional auto-regressive model; Model selection; Penalized likelihood; Simultaneous auto-regressive model; Spatial statistics; Variable selection

1. Introduction

In many fields of the biological and physical sciences, rapid advances in technical capabilities have dramatically increased the amount of data that are collected across space. Here we restrict our attention to spatial data observed on a lattice. In particular, many remotely sensed data in ecological and environmental studies are aggregated at a certain resolution on a regular grid. Spatial linear models are important tools for the analysis of such data and have been applied in a wide range of disciplines (see, for example, Cressie (1993) and Schabenberger and Gotway (2005)). In these models, linear regression is specified to associate a response variable with covariates. Because data are arranged on a spatial lattice, auto-regressive models are used to account for spatial dependence by associating the response variable at a given site with those response variables at neighbouring sites according to a neighbourhood structure. Statistical inference for the regression coefficients and the auto-regressive coefficients can be carried out via likelihood-based approaches or Bayesian hierarchical modelling. However, statistical methodology for principled selection of covariates and neighbourhood structure is limited and will be the focus of this paper.

Much research has been accomplished on variable selection in standard linear regression. Most recently, penalized methods are becoming increasingly popular. For example, Tibshirani

Address for correspondence: Jun Zhu, Department of Statistics, Colorado State University, Fort Collins, CO 80523, USA.
E-mail: jzhu@stat.colostate.edu

 1369–7412/10/72389

(1996) proposed a least absolute shrinkage and selection operator (the lasso), which performs simultaneous variable selection and parameter estimation. Zou (2006) improved the original lasso by developing an adaptive lasso, which utilizes smaller penalties for larger coefficients to remove the estimation bias, and hence it enjoys the oracle properties that can only be achieved by the original lasso under certain conditions. Efron *et al.* (2004) devised the least angle regression algorithm LARS which allows computing all lasso estimates along a path of its tuning parameter with the same computation order as the single ordinary least squares estimate based on the full model. See also Hesterberg *et al.* (2008) and the references therein for an up-to-date comprehensive review of variable selection.

Although most of these penalized methods deal with independent data, there are some results for dependent data. It is challenging to extend the penalized methods to data that are dependent either over time or across space, as variable selection involves not only regression coefficients but also auto-correlation coefficients. For example, Wang *et al.* (2007a) considered auto-regressive time series models and developed a lasso for selecting both regression coefficients and auto-regressive coefficients. For lattice data, Huang *et al.* (2010) proposed a spatial lasso, although the types of model were somewhat restricted and no theoretical properties were established. Most recently, Zhu and Liu (2009) developed estimation of spatial covariance matrices by using a penalized likelihood approach with weighted L_1-regularization. Although the form of spatial covariance matrix is non-parametric and thus flexible, the method in its current form does not address regression and the estimation procedure requires ordering of the sampling sites as well as replications of samples over time.

In general, model selection for spatial data is underdeveloped and demands further research. Owing to a lack of systematic approaches, selection of spatial linear models in practice, particularly that of a neighbourhood structure, is often not addressed or based on *ad hoc* methods with little understanding of the statistical properties. For example, a practitioner may prespecify the order of neighbourhood with or without considering covariates. Then, given the neighbourhood structure, covariates are selected. It is rare that covariates and neighbourhood structures are selected simultaneously. Furthermore, it is often computationally intensive and thus infeasible to compare and select among all possible combinations of models. Computational cost is a critical issue for spatial data analysis, as the sample size tends to be large and the dependence structure tends to be complex. Thus we believe that penalized methods accompanied by efficient computational algorithms would be especially suitable for selection of spatial linear models.

Our work here concerns, in essence, an important extension from those in Wang *et al.* (2007a) for time series data in one dimension to spatial data in multi-dimensional space. We consider spatial linear models in general, as well as conditionally and simultaneously specified auto-regressive models as two special cases. In Section 2, we propose a flexible form of parameterization of spatial dependence, which eases both practical interpretation and model selection. In Section 3, we develop a spatial adaptive lasso for selection of not only covariates but also a neighbourhood structure. An efficient algorithm is developed for computing penalized maximum likelihood estimates (MLEs). In particular, we devise a LARS-type algorithm that approximates the penalized maximum likelihood estimates of regression and auto-regressive coefficients. We establish theoretical properties of penalized maximum likelihood estimates and their approximations in terms of consistency, sparsity and asymptotic normality in Section 4. We address estimation of standard errors and regularization parameters in Section 5. Finally, in Section 6, we demonstrate, via simulation and a real data example, that our proposed methodology has sound finite sample properties and can be useful in practical applications. Technical proofs are given as appendices in supplementary material at `www.stat.sinica.edu.tw/~hchuang/paper.html`.

2. Spatial linear model

2.1. Model specification

On a spatial lattice $D_n = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\} \subset \mathbb{R}^d (d \in \mathbb{N})$, we consider response variables $\{y_i \equiv y(\mathbf{s}_i) : \mathbf{s}_i \in D_n, i = 1, \ldots, n\}$ with

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i, \tag{1}$$

where $\mathbf{x}_i = (x_{1i}, \ldots, x_{pi})'$ is a p-dimensional vector of covariates at site $\mathbf{s}_i$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a p-dimensional vector of regression coefficients. Without loss of generality we standardize the individual covariates to have mean 0 and variance 1 (Wang *et al.*, 2007a).

The errors $\{\varepsilon_i : i = 1, \ldots, n\}$ are assumed to be a zero-mean Gaussian process and

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Gamma}), \tag{2}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ denotes an n-dimensional vector of errors and $\boldsymbol{\Gamma} = [\mathrm{cov}(\varepsilon_i, \varepsilon_j)]_{i,j=1}^n$ is an $n \times n$ matrix consisting of the covariance of errors.

A special case of model (2) is a conditionally specified Gaussian model with

$$\boldsymbol{\Gamma} = (\mathbf{I}_n - \mathbf{C})^{-1} \mathbf{V}, \tag{3}$$

where $\mathbf{I}_n$ is an identity matrix, $\mathbf{C} = [c_{ij}]_{i,j=1}^n$, $\mathbf{I}_n - \mathbf{C}$ is non-singular, $\mathbf{V} = \mathrm{diag}(\{\sigma_i^2\}_{i=1}^n)$ and $(\mathbf{I}_n - \mathbf{C})^{-1}\mathbf{V}$ is symmetric and positive definite, all of dimension $n \times n$. The model is also known as the conditional auto-regressive (CAR) model, as it can be formulated by the conditional distribution of the error as

$$E(\varepsilon_i | \varepsilon_j : j \neq i) = \sum_{j=1}^n c_{ij} \varepsilon_j,$$
$$\mathrm{var}(\varepsilon_i | \varepsilon_j : j \neq i) = \sigma_i^2.$$

To ensure a valid joint distribution of $\boldsymbol{\varepsilon}$, it is necessary that $c_{ii} = 0$, $c_{ij}\sigma_j^2 = c_{ji}\sigma_i^2$ and $c_{ij} = 0$ if site $j \notin \mathcal{N}(i)$, where $\mathcal{N}(i)$ denotes the neighbourhood of site i consisting of indices of sites that are neighbours of site i according to a neighbourhood structure (Besag, 1974; Cressie, 1993).

Another special case of model (2) is a simultaneously specified Gaussian model with

$$\boldsymbol{\Gamma} = (\mathbf{I}_n - \mathbf{C})^{-1} \mathbf{V} (\mathbf{I}_n - \mathbf{C}')^{-1}, \tag{4}$$

where $\mathbf{C} = [c_{ij}]_{i,j=1}^n$, $\mathbf{I}_n - \mathbf{C}$ is non-singular and $\mathbf{V} = \mathrm{diag}(\{\sigma_i^2\}_{i=1}^n)$. The model is also known as the simultaneous auto-regressive (SAR) model, as it can be formulated by

$$\boldsymbol{\varepsilon} = \mathbf{C}\boldsymbol{\varepsilon} + \boldsymbol{\nu},$$

where $\boldsymbol{\nu} \sim N(\mathbf{0}, \mathbf{V})$ denotes an n-dimensional vector of independent noise (Cressie, 1993).

2.2. Neighbourhood structure parameterization

For a given site i, let $\mathcal{N}(i) = \cup_{k=1}^q \mathcal{N}_k(i)$, where $\{\mathcal{N}_k(i) : k = 1, \ldots, q\}$ are neighbourhoods that partition $\mathcal{N}(i)$, $i = 1, \ldots, n$. We propose a general class of $\mathbf{C}$ in the form of

$$\mathbf{C} = \sum_{k=1}^q \theta_k \mathbf{W}_k, \tag{5}$$

where θ_k is an auto-regressive coefficient and $\mathbf{W}_k = [w_{ij}^k]_{i,j=1}^n$ is an $n \times n$ matrix consisting of spatial weights. The partition of $\mathcal{N}(i)$ is flexible. Here we focus on a regular grid. We define the kth-order neighbours in $\mathcal{N}_k(i)$ of a given site i as the kth-nearest neighbours in terms of distance

between two sites, for $k = 1, \ldots, q$. Thus, $N_1(i)$ consists of the four nearest neighbours in the north, south, west and east, $N_2(i)$ consists of the four second-nearest neighbours in the north-west, north-east, south-west and south-east, etc. The number of neighbours is not necessarily 4 at higher orders. To accommodate anisotropy, we could further partition $\mathcal{N}_k(i)$ according to direction. For instance, we may let $N_{1,\updownarrow}(i)$ consist of the two nearest neighbours in the north and south direction whereas $N_{1,\leftrightarrow}(i)$ consist of the two nearest neighbours in the west and east direction. In general, the magnitude of $\{\theta_k\}$ reflects not only the extent but also the direction of spatial auto-correlation across space.

For a CAR model, sufficient conditions to ensure a valid joint distribution of the errors are $w_{ii}^k = 0$, $w_{ij}^k \sigma_j^2 = w_{ji}^k \sigma_i^2$ and $w_{ij}^k = 0$ if site $j \notin \mathcal{N}_k(i)$. We further assume that $w_{ij}^k = w_{ji}^k$, for all $i, j = 1, \ldots, n$ and $k = 1, \ldots, q$. Thus, $\mathbf{C}$ is symmetric and $\sigma_i^2 \equiv \sigma^2$ with

$$\mathbf{V} = \sigma^2 \mathbf{I}_n. \tag{6}$$

For an SAR model, although not necessary, we assume that $w_{ij}^k = w_{ji}^k$, for all $i, j = 1, \ldots, n$, $k = 1, \ldots, q$ and $\sigma_i^2 \equiv \sigma^2$.

3. Model selection

3.1. Spatial adaptive lasso

Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)'$ denote a q-dimensional vector of auto-regressive coefficients and $\boldsymbol{\gamma} = (\boldsymbol{\theta}', \sigma^2)'$. We sometimes use $\boldsymbol{\Gamma}_\gamma$ and $\mathbf{C}_\theta$ to emphasize the parameterization of $\boldsymbol{\Gamma}$ and $\mathbf{C}$ by $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$. Let $\mathbf{y} = (y_1, \ldots, y_n)'$ denote an n-dimensional vector of response variables and let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ denote an $n \times p$ design matrix, where $\mathbf{x}_j = (x_{j1}, \ldots, x_{jn})'$ denotes an n-dimensional vector of the jth covariate with $j = 1, \ldots, p$. Thus, by expressions (1) and (2),

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Gamma}_\gamma). \tag{7}$$

We consider selection of covariates and neighbourhood orders by determining which regression coefficients and which auto-regressive coefficients are non-zero and then estimate the non-zero coefficients. Our proposed method enables simultaneous model selection and parameter estimation.

Let $\boldsymbol{\eta} = (\boldsymbol{\beta}', \boldsymbol{\gamma}')'$ denote a $(p+q+1)$-dimensional vector of model parameters consisting of both regression coefficients and auto-regressive coefficients. Under model (7), the log-likelihood function is

$$\begin{aligned} \log\{L(\boldsymbol{\eta}; \mathbf{y}, \mathbf{X})\} &= \text{constant} - \tfrac{1}{2} \log |\boldsymbol{\Gamma}_\gamma| - \tfrac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \boldsymbol{\Gamma}_\gamma^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &\equiv \text{constant} + l(\boldsymbol{\eta}). \end{aligned} \tag{8}$$

We let

$$\hat{\boldsymbol{\eta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\eta}} \{l(\boldsymbol{\eta})\}$$

denote the MLEs of $\boldsymbol{\eta}$.

We consider the penalized log-likelihood function

$$\begin{aligned} Q(\boldsymbol{\eta}) &= l(\boldsymbol{\eta}) - n \sum_{j=1}^{p} \lambda_j |\beta_j| - n \sum_{k=1}^{q} \tau_k |\theta_k| \\ &= -\tfrac{1}{2} \log |\boldsymbol{\Gamma}_\gamma| - \tfrac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \boldsymbol{\Gamma}_\gamma^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - n \sum_{j=1}^{p} \lambda_j |\beta_j| - n \sum_{k=1}^{q} \tau_k |\theta_k|, \end{aligned} \tag{9}$$

where the last two terms are the adaptive lasso penalty on the coefficients, $\{\lambda_j\}_{j=1}^p$ are regularization parameters for the regression coefficients $\boldsymbol{\beta}$ and $\{\tau_k\}_{k=1}^q$ are regularization parameters

for the auto-regressive coefficients $\boldsymbol{\theta}$. We let

$$\hat{\boldsymbol{\eta}}_{\mathrm{PMLE}} = \arg\max_{\boldsymbol{\eta}} \{Q(\boldsymbol{\eta})\}$$

denote the penalized maximum likelihood estimates (PMLEs) of $\boldsymbol{\eta}$.

For a CAR model with expressions (3), (5) and (6), the penalized log-likelihood function (9) becomes

$$\begin{aligned} Q(\boldsymbol{\eta}) = &-\frac{n}{2}\log(\sigma^2) + \frac{1}{2}\log|\mathbf{I}_n - \mathbf{C}_{\boldsymbol{\theta}}| - \frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{I}_n - \mathbf{C}_{\boldsymbol{\theta}})(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\ &- n\sum_{j=1}^{p}\lambda_j|\beta_j| - n\sum_{k=1}^{q}\tau_k|\theta_k|. \end{aligned}$$

For an SAR model with expressions (4), (5) and (6), the penalized log-likelihood function (9) becomes

$$\begin{aligned} Q(\boldsymbol{\eta}) = &-\frac{n}{2}\log(\sigma^2) + \frac{1}{2}\log|(\mathbf{I}_n - \mathbf{C}'_{\boldsymbol{\theta}})(\mathbf{I}_n - \mathbf{C}_{\boldsymbol{\theta}})| - \frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{I}_n - \mathbf{C}'_{\boldsymbol{\theta}})(\mathbf{I}_n - \mathbf{C}_{\boldsymbol{\theta}})(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\ &- n\sum_{j=1}^{p}\lambda_j|\beta_j| - n\sum_{k=1}^{q}\tau_k|\theta_k|. \end{aligned}$$

3.2. Penalized maximum likelihood via LARS$_m$

Let $\hat{\boldsymbol{\eta}}^{(0)} = (\hat{\boldsymbol{\beta}}^{(0)'}, \hat{\boldsymbol{\gamma}}^{(0)'})'$ denote an initial value of $\boldsymbol{\eta}$, which is set to the MLE $\hat{\boldsymbol{\eta}}_{\mathrm{MLE}}$ at the beginning of the iterations. At iteration $m = 1, 2, \ldots$, when $\boldsymbol{\eta} \approx \hat{\boldsymbol{\eta}}^{(m-1)}$, we approximate the penalized log-likelihood function (9) up to a constant by

$$\begin{aligned} Q^*(\boldsymbol{\eta}) = &(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}^{(m-1)})' \frac{\partial l(\hat{\boldsymbol{\eta}}^{(m-1)})}{\partial \boldsymbol{\eta}} - \frac{1}{2}(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}^{(m-1)})' \mathcal{I}(\hat{\boldsymbol{\eta}}^{(m-1)})(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}^{(m-1)}) \\ &- n\sum_{j=1}^{p}\lambda_j|\beta_j| - n\sum_{k=1}^{q}\tau_k|\theta_k|, \end{aligned} \tag{10}$$

where

$$\mathcal{I}(\boldsymbol{\eta}) = E_{\eta}\left\{ -\frac{\partial^2 l(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}\partial \boldsymbol{\eta}'} \right\}$$

is an information matrix and the regularization parameters are separate for each iteration. We propose to update $\hat{\boldsymbol{\eta}}^{(m-1)}$ to

$$\hat{\boldsymbol{\eta}}^{(m)} = \arg\max_{\boldsymbol{\eta}} \{Q^*(\boldsymbol{\eta})\}, \tag{11}$$

and to iterate equation (11) until convergence.

Since the information matrix is a block diagonal matrix $\mathcal{I}(\boldsymbol{\eta}) = \mathrm{diag}\{\mathcal{I}(\boldsymbol{\beta}), \mathcal{I}(\boldsymbol{\gamma})\}$ (see Section 5), we update $\hat{\boldsymbol{\beta}}^{(m-1)}$ and $\hat{\boldsymbol{\gamma}}^{(m-1)}$ separately, i.e. we update $\hat{\boldsymbol{\beta}}^{(m-1)}$ to

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{(m)} = \arg\min_{\boldsymbol{\beta}} \Bigg\{ &-(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(m-1)})' \frac{\partial l(\hat{\boldsymbol{\eta}}^{(m-1)})}{\partial \boldsymbol{\beta}} + \frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(m-1)})' \mathcal{I}(\hat{\boldsymbol{\beta}}^{(m-1)})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(m-1)}) \\ &+ n\sum_{j=1}^{p}\lambda_j|\beta_j| \Bigg\}. \end{aligned} \tag{12}$$

It can be shown that the solution of equation (12) can be attained equivalently by

$$\hat{\boldsymbol{\beta}}^{*(m)} = \arg\min_{\boldsymbol{\beta}^*} \left\{ \tfrac{1}{2}(\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}^*)'(\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}^*) + n \sum_{j=1}^{p} |\beta_j^*| \right\}, \tag{13}$$

where

$$\mathbf{y}^* = (\mathbf{A}^{-1})' \left\{ \frac{\partial l(\hat{\boldsymbol{\eta}}^{(m-1)})}{\partial \boldsymbol{\beta}} + \mathcal{I}(\hat{\boldsymbol{\beta}}^{(m-1)})\hat{\boldsymbol{\beta}}^{(m-1)} \right\},$$

$\mathbf{X}^* = \mathbf{A}\,\mathrm{diag}(\{\lambda_j^{-1}\}_{j=1}^p)$, $\boldsymbol{\beta}^* = \mathrm{diag}(\{\lambda_j\}_{j=1}^p)\boldsymbol{\beta}$ and $\mathcal{I}(\hat{\boldsymbol{\beta}}^{(m-1)}) = \mathbf{A}'\mathbf{A}$. Hence, $\hat{\boldsymbol{\beta}}^{(m)} = \mathrm{diag}(\{\lambda_j^{-1}\}_{j=1}^p)\hat{\boldsymbol{\beta}}^{*(m)}$.

Similarly, we update $\hat{\boldsymbol{\gamma}}^{(m-1)}$ to

$$\hat{\boldsymbol{\gamma}}^{(m)} = \arg\min_{\boldsymbol{\gamma}} \left\{ -(\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}^{(m-1)})' \frac{\partial l(\hat{\boldsymbol{\eta}}^{(m-1)})}{\partial \boldsymbol{\gamma}} + \frac{1}{2}(\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}^{(m-1)})' \mathcal{I}(\hat{\boldsymbol{\gamma}}^{(m-1)})(\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}^{(m-1)}) + n \sum_{k=1}^{q} \tau_k |\theta_k| \right\}. \tag{14}$$

Since σ^2 is not subject to any penalty, we update the terms differently. We let

$$\mathbf{X}_k^{**} = \tau_k^{-1}(\mathbf{B}_k - c_k\mathbf{B}_{q+1}), \qquad k = 1, \ldots, q,$$
$$\mathbf{X}_{q+1}^{**} = \mathbf{B}_{q+1},$$

where $c_k = \mathbf{B}_{q+1}'\mathbf{B}_k / \mathbf{B}_{q+1}'\mathbf{B}_{q+1}$, for $k = 1, \ldots, q$, and $\mathcal{I}(\hat{\boldsymbol{\gamma}}^{(m-1)}) = \mathbf{B}'\mathbf{B}$. It is obvious that $\mathbf{X}_{q+1}^{**\prime}\mathbf{X}_k^{**} = \mathbf{0}$ for $k = 1, \ldots, q$. Let

$$\mathbf{y}^{**} = (\mathbf{B}^{-1})' \left\{ \frac{\partial l(\hat{\boldsymbol{\eta}}^{(m-1)})}{\partial \boldsymbol{\gamma}} + \mathcal{I}(\hat{\boldsymbol{\gamma}}^{(m-1)})\hat{\boldsymbol{\gamma}}^{(m-1)} \right\}.$$

It can be shown that the solution of σ^2 in equation (14) can be attained in closed form as

$$(\hat{\sigma}^{*2})^{(m)} = \mathbf{X}_{q+1}^{**\prime}\mathbf{y}^{**} / \mathbf{X}_{q+1}^{**\prime}\mathbf{X}_{q+1}^{**},$$

where $\sigma^{*2} = \sum_{k=1}^{q} c_k\theta_k + \sigma^2$. Further,

$$\hat{\boldsymbol{\theta}}^{*(m)} = \arg\min_{\boldsymbol{\theta}^*} \left\{ \tfrac{1}{2}(\mathbf{y}^{**} - \mathbf{X}^{**}\boldsymbol{\theta}^*)'(\mathbf{y}^{**} - \mathbf{X}^{**}\boldsymbol{\theta}^*) + n \sum_{k=1}^{q} |\theta_k^*| \right\}, \tag{15}$$

where $\mathbf{X}^{**} = (\mathbf{X}_1^{**}, \ldots, \mathbf{X}_q^{**})$ and $\boldsymbol{\theta}^* = \mathrm{diag}(\{\tau_k\}_{k=1}^q)\boldsymbol{\theta}$. Hence, $\hat{\boldsymbol{\theta}}^{(m)} = \mathrm{diag}(\{\tau_k^{-1}\}_{k=1}^q)\hat{\boldsymbol{\theta}}^{*(m)}$ and

$$(\hat{\sigma}^2)^{(m)} = (\hat{\sigma}^{*2})^{(m)} - \sum_{k=1}^{q} c_k\hat{\theta}_k^{(m)}.$$

At convergence, we let $\hat{\boldsymbol{\eta}}_{\mathrm{APMLE}}$ denote the approximate penalized maximum likelihood estimates (APMLEs) of $\boldsymbol{\eta}$. Within each iteration, equations (13) and (15) can be solved by a LARS algorithm and, thus, the computation is highly efficient. Our proposed algorithm henceforth will be referred to as a multiple-step LARS algorithm $\mathrm{LARS}_{\mathrm{m}}$.

4. Asymptotic properties

4.1. Notation

Let $\gamma^0 = (\theta_1^0, \ldots, \theta_q^0, (\sigma^2)^0)' = (\boldsymbol{\theta}_1^{0'}, \boldsymbol{\theta}_2^{0'}, (\sigma^2)^0)'$ denote a $(q+1)$-dimensional vector of true parameter values, where without loss of generality we assume that $\boldsymbol{\theta}_1^0$ is a t-dimensional vector of non-zero auto-regressive coefficients and $\boldsymbol{\theta}_2^0 = \mathbf{0}$ is a $(q-t)$-dimensional vector of zero-valued auto-regressive coefficients. In general, we shall write $\gamma = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2', \sigma^2)'$ and its estimate as $\hat{\gamma} = (\hat{\boldsymbol{\theta}}_1', \hat{\boldsymbol{\theta}}_2', \hat{\sigma}^2)'$. We denote $\boldsymbol{\beta}$, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, and the corresponding true values and estimates in a similar manner but replacing q with p, and t with s, and leaving out σ^2. Let $\eta^0 = (\boldsymbol{\beta}_1^{0'}, \boldsymbol{\beta}_2^{0'}, \boldsymbol{\theta}_1^{0'}, \boldsymbol{\theta}_2^{0'}, (\sigma^2)^0)'$ denote a $(p+q+1)$-dimensional vector of true parameter values. Let $\eta_1^0 = (\boldsymbol{\beta}_1^{0'}, \boldsymbol{\theta}_1^{0'}, (\sigma^2)^0)'$ denote an $(s+t+1)$-dimensional vector of non-zero parameters and $\eta_2^0 = (\boldsymbol{\beta}_2^{0'}, \boldsymbol{\theta}_2^{0'})'$ a $(p+q-s-t)$-dimensional vector of zero-valued parameters. Also, let $\gamma_1^0 = (\boldsymbol{\theta}_1^{0'}, (\sigma^2)^0)'$. Then η, η_1, η_2 and γ_1 and the corresponding estimates are defined similarly. Let $a_n = \max\{\lambda_j : j = 1, \ldots, s, \tau_k : k = 1, \ldots, t\}$ and $b_n = \min\{\lambda_j : j = s+1, \ldots, p, \tau_k : k = t+1, \ldots, q\}$. Note that a_n and b_n are associated with the regularization parameters in the penalty.

4.2. Asymptotic properties

Under suitable regularity conditions (A.1)–(A.4) (see Appendix A of the web-based supplementary materials), we establish the oracle properties of the PMLE $\hat{\eta}_{\mathrm{PMLE}}$.

Theorem 1. Suppose that conditions (A.1)–(A.4) hold and $a_n = O(n^{-1/2})$ as $n \to \infty$.

(a) With probability tending to 1, there is a local maximizer $\hat{\eta}$ of $Q(\eta)$ defined in equation (9) such that $\|\hat{\eta} - \eta^0\| = O_p(n^{-1/2} + a_n)$.
(b) If, in addition, $n^{1/2} b_n \to \infty$ as $n \to \infty$, then, with probability tending to 1, $\hat{\eta}_2 = \mathbf{0}$, i.e. $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$ and $\hat{\boldsymbol{\theta}}_2 = \mathbf{0}$.
(c) If, in addition, $a_n = o(n^{-1/2})$, then

$$n^{1/2}(\hat{\eta}_1 - \eta_1^0) \xrightarrow{\mathrm{D}} N\{\mathbf{0}, \mathbf{J}(\eta_1^0)^{-1}\},$$

where

$$\mathbf{J}(\eta_1^0) = \begin{pmatrix} \mathbf{J}(\boldsymbol{\beta}_1^0) & \mathbf{0} \\ \mathbf{0} & \mathbf{J}(\gamma_1^0) \end{pmatrix},$$

$\mathbf{J}(\boldsymbol{\beta}_1^0)$ consists of the first $s \times s$ upper left submatrix of $\mathbf{J}(\boldsymbol{\beta}^0)$, $\mathbf{J}(\gamma_1^0)$ consists of the submatrix of $\mathbf{J}(\gamma^0)$ corresponding to rows $1, \ldots, t$ and $q+1$, and $\mathbf{J}(\boldsymbol{\beta}^0)$ and $\mathbf{J}(\gamma^0)$ are defined in condition (A.4).

Theorem 1, part (a), establishes the existence of the PMLE as well as consistency at the rate of $n^{1/2}$. Theorem 1, part (b), ensures sparsity of the PMLE, i.e., as $n \to \infty$, the PMLEs of zero-valued regression coefficients and zero-valued auto-regressive coefficients are 0, with probability tending to 1. Theorem 1, part (c), is a central limit theorem for the PMLE of the non-zero-valued regression and auto-regressive coefficients. On the basis of the asymptotic variance in the limiting normal distribution, we can approximate the variance of $\hat{\boldsymbol{\beta}}_1$ by $\mathcal{I}(\boldsymbol{\beta}_1^0)^{-1}$ and that of $\hat{\gamma}_1$ by $\mathcal{I}(\gamma_1^0)^{-1}$.

We then establish the oracle properties of $\hat{\eta}_{\mathrm{APMLE}} = (\hat{\eta}'_{\mathrm{APMLE},1}, \hat{\eta}'_{\mathrm{APMLE},2})'$, which is an approximation obtained by the $\mathrm{LARS_m}$ algorithm that was devised in Section 3.2.

Theorem 2. Suppose that conditions (A.1)–(A.4) hold, $n^{1/2} b_n \to \infty$ as $n \to \infty$ and $a_n = o(n^{-1/2})$. Then, with probability tending to 1, $\hat{\eta}_2^{(m)} = \mathbf{0}$, and

$$n^{1/2}(\hat{\eta}_1^{(m)} - \eta_1^0) \xrightarrow{\mathrm{D}} N\{\mathbf{0}, \mathbf{J}(\eta_1^0)^{-1}\},$$

for all $m \in \mathbb{N} \equiv \{1, 2, \ldots\}$, where $\hat{\boldsymbol{\eta}}^{(m)} = (\hat{\boldsymbol{\eta}}_1^{(m)'}, \hat{\boldsymbol{\eta}}_2^{(m)'})'$. In particular, with probability tending to 1, $\hat{\boldsymbol{\eta}}_{\mathrm{APMLE},2} = \mathbf{0}$, and

$$n^{1/2}(\hat{\boldsymbol{\eta}}_{\mathrm{APMLE},1} - \boldsymbol{\eta}_1^0) \xrightarrow{\mathrm{D}} N\{\mathbf{0}, \mathbf{J}(\boldsymbol{\eta}_1^0)^{-1}\}.$$

The proofs of theorems 1 and 2 are shown in appendix B and appendix C in the Web-based supplementary materials. Although $\hat{\boldsymbol{\eta}}_{\mathrm{APMLE}}$ obtained from our iterative computational algorithm is a local optimum and is not necessarily a global optimum, theorem 2 shows that it has desirable asymptotic properties including sparsity and the oracle property. Moreover, these asymptotic properties hold for each step of the iteration and, thus, convergence is of less concern than those algorithms such that the asymptotic properties of only the solution at convergence are known.

5. Further computational aspects

The standard errors of the APMLE can be estimated according to theorem 2. Let $\boldsymbol{\Gamma}^k = \partial \boldsymbol{\Gamma}^{-1}/\partial \gamma_k$, $\boldsymbol{\Gamma}_k = \partial \boldsymbol{\Gamma}/\partial \gamma_k = -\boldsymbol{\Gamma}\boldsymbol{\Gamma}^k\boldsymbol{\Gamma}$, $\boldsymbol{\Gamma}^{kk'} = \partial^2 \boldsymbol{\Gamma}^{-1}/\partial \gamma_k \partial \gamma_{k'}$ and

$$\boldsymbol{\Gamma}_{kk'} = \frac{\partial^2 \boldsymbol{\Gamma}}{\partial \gamma_k \partial \gamma_{k'}} = \boldsymbol{\Gamma}(\boldsymbol{\Gamma}^k \boldsymbol{\Gamma} \boldsymbol{\Gamma}^{k'} + \boldsymbol{\Gamma}^{k'} \boldsymbol{\Gamma} \boldsymbol{\Gamma}^k - \boldsymbol{\Gamma}^{kk'})\boldsymbol{\Gamma},$$

for $k, k' = 1, \ldots, q+1$ where, for ease of presentation, $\gamma_{q+1} = \sigma^2$. From equation (8), we have

$$\begin{aligned}
\frac{\partial l(\boldsymbol{\eta})}{\partial \boldsymbol{\beta}} &= \mathbf{X}'\boldsymbol{\Gamma}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \\
\frac{\partial l(\boldsymbol{\eta})}{\partial \gamma_k} &= -\frac{1}{2}\mathrm{tr}(\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_k) - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Gamma}^k(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\
&= \frac{1}{2}\mathrm{tr}(\boldsymbol{\Gamma}^k\boldsymbol{\Gamma}) - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Gamma}^k(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).
\end{aligned}$$

Moreover, we have

$$\begin{aligned}
\frac{\partial^2 l(\boldsymbol{\eta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} &= -\mathbf{X}'\boldsymbol{\Gamma}^{-1}\mathbf{X}, \\
\frac{\partial^2 l(\boldsymbol{\eta})}{\partial \boldsymbol{\beta} \partial \gamma_k} &= \mathbf{X}'\boldsymbol{\Gamma}^k(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \\
\frac{\partial^2 l(\boldsymbol{\eta})}{\partial \gamma_k \partial \gamma_{k'}} &= -\frac{1}{2}\mathrm{tr}(\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_{kk'} + \boldsymbol{\Gamma}^{k'}\boldsymbol{\Gamma}_k) - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Gamma}^{kk'}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\
&= -\frac{1}{2}\mathrm{tr}(\boldsymbol{\Gamma}^k\boldsymbol{\Gamma}\boldsymbol{\Gamma}^{k'}\boldsymbol{\Gamma} - \boldsymbol{\Gamma}^{kk'}\boldsymbol{\Gamma}) - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Gamma}^{kk'}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).
\end{aligned}$$

Since $E_{\boldsymbol{\eta}}\{-\partial^2 l(\boldsymbol{\eta})/\partial \boldsymbol{\beta}\, \partial \boldsymbol{\gamma}'\} = \mathbf{0}$, the information matrix of $\mathbf{y}$ is

$$\mathcal{I}(\boldsymbol{\eta}) = \mathrm{diag}\{\mathcal{I}(\boldsymbol{\beta}), \mathcal{I}(\boldsymbol{\gamma})\}, \tag{16}$$

where

$$\mathcal{I}(\boldsymbol{\beta}) = E_{\boldsymbol{\eta}}\left\{-\frac{\partial^2 l(\boldsymbol{\eta})}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}'}\right\} = \mathbf{X}'\boldsymbol{\Gamma}^{-1}\mathbf{X},$$

and the (k, k')th entry of

$$\mathcal{I}(\gamma) = E_{\eta}\left\{ -\frac{\partial^2 l(\boldsymbol{\eta})}{\partial\gamma\,\partial\gamma'} \right\}$$

is $\frac{1}{2}\mathrm{tr}(\boldsymbol{\Gamma}^k\boldsymbol{\Gamma}\boldsymbol{\Gamma}^{k'}\boldsymbol{\Gamma})$.

In particular, we note that,

$$\begin{aligned} \mathrm{var}(\hat{\boldsymbol{\beta}}_1) &\approx \mathcal{I}(\boldsymbol{\beta}_1)^{-1}, \\ \mathrm{var}(\hat{\boldsymbol{\theta}}_1) &\approx \mathcal{I}(\gamma_1)^{-1}, \end{aligned} \tag{17}$$

where $\mathcal{I}(\boldsymbol{\beta}_1)$ is the first $s \times s$ upper left submatrix of $\mathcal{I}(\boldsymbol{\beta}) = \mathbf{X}'\boldsymbol{\Gamma}^{-1}\mathbf{X}$, and $\mathcal{I}(\gamma_1)$ is the submatrix of $\mathcal{I}(\gamma)$ corresponding to rows $1, \ldots, t$, and $q+1$. Evaluating expression (17) at the APMLE, we obtain estimates of the corresponding variance–covariance matrices.

Finally, to estimate the regularization parameters $\{\lambda_j\}_{j=1}^{p}$ and $\{\tau_k\}_{k=1}^{q}$, we let

$$\begin{aligned} \lambda_j &= \lambda \log(n)(n|\hat{\beta}_j|)^{-1}, \\ \tau_k &= \tau \log(n)(n|\hat{\theta}_k|)^{-1}. \end{aligned} \tag{18}$$

The dimension reduction in expression (18) is useful, as now only two regularization parameters instead of $p+q$ need to be estimated (Wang *et al.*, 2007a). To determine λ and τ, we compute the Bayesian information criterion BIC,

$$\mathrm{BIC}(\lambda, \tau) = -2\, l(\hat{\boldsymbol{\eta}}; \lambda, \tau) + e(\lambda, \tau)\log(n), \tag{19}$$

where

$$e(\lambda, \tau) = \sum_{j=1}^{p} I\{\hat{\beta}_j \neq 0\} + \sum_{k=1}^{q} I\{\hat{\theta}_k \neq 0\},$$

for all combinations of λ and τ (Wang *et al.*, 2007b). In each iteration, we select the combination that has the smallest BIC-values. Since we utilize a LARS-type algorithm which is a path algorithm, we may obtain the best λ and τ in a computationally efficient manner. To reduce the dimension of the regularization parameters further, we consider one regularization parameter, as suggested by a referee. The computational procedure is similar to that for two regularization parameters, but there may be a computational advantage due to the additional dimension reduction.

6. Numerical examples

6.1. Simulation study

For simulation, we consider $m \times m$ square lattices, where $m = 5, 10, 15$. The corresponding sample sizes are $n = 25, 100, 225$. For regression, there are seven covariates following a standard normal distribution. The cross-covariate correlation is assumed to be $\mathrm{corr}(x_j, x_{j'}) = \rho^{|j-j'|}$, where $\rho = 0.5$ and $j, j' = 1, \ldots, 7$, whereas each individual covariate is assumed to be dependent across space and to have an exponential covariance function with no nugget and range parameter 1. The regression coefficients are set to $\boldsymbol{\beta} = (4, 3, 2, 1, 0, 0, 0)'$. We let the error term have mean 0, following either an SAR model or a CAR model. The neighbourhood structure is from the first to the fifth order, where the first-order neighbourhood consists of the nearest neighbours, the second-order neighbourhood consists of the second-nearest neighbours, and so on, of a given site. The auto-regressive coefficients are set to $\boldsymbol{\theta} = (0.2, 0, 0, 0, 0)'$. For each m, a total of 100 data sets are simulated according to the spatial linear model (1)–(2).

For comparison, we consider three alternatives. The first alternative is a simplification of our $\mathrm{LARS_m}$ algorithm. Rather than iterating until convergence, the one-step LARS algorithm $\mathrm{LARS_1}$ stops after just one iteration. $\mathrm{LARS_1}$ is, in spirit, similar to the one-step-ahead local linear approximation that was developed by Zou and Li (2008) for regression models with independent errors. For both $\mathrm{LARS_m}$ and $\mathrm{LARS_1}$, we consider one and two regularization parameters.

The second alternative is a local quadratic approximation (LQA) algorithm that was proposed by Fan and Li (2001). Although developed originally for regression with independent errors, Wang *et al.* (2007b) further developed the LQA for time series. For the spatial setting under consideration here, we need first to extend this algorithm. The extended LQA algorithm is based on the approximation

$$\begin{aligned} Q(\boldsymbol{\eta}) \approx \text{constant} &+ (\boldsymbol{\eta}-\tilde{\boldsymbol{\eta}}^{(0)})' \frac{\partial l(\tilde{\boldsymbol{\eta}}^{(0)})}{\partial \boldsymbol{\eta}} - \frac{1}{2}(\boldsymbol{\eta}-\tilde{\boldsymbol{\eta}}^{(0)})' \mathcal{I}(\tilde{\boldsymbol{\eta}}^{(0)})(\boldsymbol{\eta}-\tilde{\boldsymbol{\eta}}^{(0)}) \\ &- \frac{1}{2} n \boldsymbol{\beta}' \boldsymbol{\Sigma}(\tilde{\boldsymbol{\beta}}^{(0)})\boldsymbol{\beta} - \frac{1}{2} n \boldsymbol{\gamma}' \boldsymbol{\Sigma}(\tilde{\boldsymbol{\gamma}}^{(0)})\boldsymbol{\gamma}, \end{aligned}$$

where $\boldsymbol{\Sigma}(\tilde{\boldsymbol{\beta}}^{(0)}) = \mathrm{diag}(\lambda_1/|\tilde{\beta}_1^{(0)}|, \ldots, \lambda_p/|\tilde{\beta}_p^{(0)}|)$ and $\boldsymbol{\Sigma}(\tilde{\boldsymbol{\gamma}}^{(0)}) = \mathrm{diag}(\tau_1/|\tilde{\theta}_1^{(0)}|, \ldots, \tau_q/|\tilde{\theta}_q^{(0)}|, 0)$.

By a Newton–Raphson step, we update $\tilde{\boldsymbol{\beta}}^{(0)}$ by

$$\begin{aligned} \tilde{\boldsymbol{\beta}}^{(1)} &= \tilde{\boldsymbol{\beta}}^{(0)} + \{\mathcal{I}(\tilde{\boldsymbol{\beta}}^{(0)}) + n\, \boldsymbol{\Sigma}(\tilde{\boldsymbol{\beta}}^{(0)})\}^{-1} \left\{ \frac{\partial l(\tilde{\boldsymbol{\eta}}^{(0)})}{\partial \boldsymbol{\beta}} - n\, \boldsymbol{\Sigma}(\tilde{\boldsymbol{\beta}}^{(0)})\tilde{\boldsymbol{\beta}}^{(0)} \right\} \\ &= \{\mathbf{X}'\boldsymbol{\Gamma}_{\tilde{\boldsymbol{\gamma}}^{(0)}}^{-1}\mathbf{X} + n\, \boldsymbol{\Sigma}(\tilde{\boldsymbol{\beta}}^{(0)})\}^{-1}\mathbf{X}'\boldsymbol{\Gamma}_{\tilde{\boldsymbol{\gamma}}^{(0)}}^{-1}\mathbf{y}, \end{aligned}$$

and update $\tilde{\boldsymbol{\gamma}}^{(0)}$ by

$$\tilde{\boldsymbol{\gamma}}^{(1)} = \tilde{\boldsymbol{\gamma}}^{(0)} + \{\mathcal{I}(\tilde{\boldsymbol{\gamma}}^{(0)}) + n\, \boldsymbol{\Sigma}(\tilde{\boldsymbol{\gamma}}^{(0)})\}^{-1} \left\{ \frac{\partial l(\tilde{\boldsymbol{\eta}}^{(0)})}{\partial \boldsymbol{\gamma}} - n\, \boldsymbol{\Sigma}(\tilde{\boldsymbol{\gamma}}^{(0)})\tilde{\boldsymbol{\gamma}}^{(0)} \right\}.$$

Then, for a small threshold value $\delta > 0$, let $\tilde{\beta}_j^{(1)} = 0$ if $|\tilde{\beta}_j^{(1)}| < \delta$ and $\tilde{\theta}_k^{(1)} = 0$ if $|\tilde{\theta}_k^{(1)}| < \delta$. The initial parameter values are set to the MLEs and the iterations continue until convergence. The selection of regularization parameters is based on BIC. Although LQA is relatively slow and once a coefficient has been shrunk to 0 it remains 0 throughout the remainder of the iterations, it has been used and shown to produce reliable results in practice.

The third alternative, which was suggested by another referee, is to use the adaptive lasso to select only the covariates, for any given possible neighbourhood structure. There are a total of 2^q possibilities, as both the size and the composition of a neighbourhood structure are of interest. Then to determine the best neighbourhood structure, we use BIC. We shall refer to this alternative as the regression-only case.

For each simulated data set, the spatial adaptive lasso was implemented using our $\mathrm{LARS_m}$ algorithm, as well as the three alternatives. Tables 1 and 2 provide the results of variable selection for CAR and SAR models respectively, in terms of the average numbers of correctly identified zero-valued and non-zero β_j, as well as zero-valued and non-zero θ_k. As the sample size increases, the number of correctly identified zero-valued (or non-zero-valued) coefficients tends to the true number. The number of correctly identified non-zero values is closer to the truth than the number of correctly identified zero values.

With few exceptions, the results of $\mathrm{LARS_m}$ are better than the first alternative $\mathrm{LARS_1}$, in terms of accuracy in variable selection. The numbers of correctly selecting non-zero-valued coefficients are comparable under one and two regularization parameters, for both $\mathrm{LARS_m}$

Table 1. Average number of correctly identified zero and non-zero regression coefficients $\{\beta_j\}$ and auto-regressive coefficients $\{\theta_k\}$ by using the proposed LARS_m method with either one or two regularization parameters and compared with three alternatives: LARS_1 with either one or two regularization parameters, the LQA and selection of regression coefficients only†

Method	*n*	*Number of non-zero* β_j	*Number of zero* β_j	*Number of non-zero* θ_k	*Number of zero* θ_k
LARS_m,	25	4.00	1.62	0.56	1.65
2-regular	100	4.00	2.48	0.97	3.66
	225	4.00	2.62	1.00	3.90
LARS_m,	25	4.00	1.77	1.00	0.83
1-regular	100	4.00	2.36	0.98	1.56
	225	4.00	2.49	1.00	1.34
LARS_1,	25	4.00	1.16	0.58	1.54
2-regular	100	4.00	1.63	0.76	3.37
	225	4.00	1.76	0.98	3.30
LARS_1,	25	4.00	1.04	0.89	0.48
1-regular	100	4.00	1.70	1.00	0.89
	225	4.00	1.77	1.00	0.83
LQA	25	3.99	1.74	0.60	2.31
	100	4.00	2.60	0.92	3.22
	225	4.00	2.59	1.00	3.50
LARS_m,	25	3.99	2.21	0.52	2.92
regression only	100	4.00	2.60	0.91	3.77
	225	4.00	2.62	0.99	3.76
LARS_1,	25	3.99	1.54	0.50	2.59
regression only	100	4.00	1.78	0.88	3.73
	225	4.00	1.85	1.00	3.90
Truth		4	3	1	4

†The sample sizes are $n = 25, 100, 225$ and the model is CAR.

and LARS_1. However, there is a tendency to miss zero-valued coefficients, especially zero-valued θ_ks when using only one regularization parameter. When compared against the second alternative LQA and the third alternative with regression coefficients only, LARS_m tends to outperform in variable selection for larger sample sizes. As for computational speed, LARS_m is slower than LARS_1, but much faster than both LQA and the regression-only case (of the order of 15–20-fold).

Additional tables are given in the Web-based supplementary materials that feature the means and standard deviations of the PMLEs of the model parameters. As the sample size increases, estimation of all the coefficients improves in terms of both accuracy and precision. The results are similar among the various methods and algorithms, under both SAR and CAR models.

6.2. Data example

A motivating example is the study of outbreaks of mountain pine beetle (MPB) *Dendroctonus ponderosae* Hopkins in western Canada. The MPB is an eruptive insect that colonizes mature pine trees via pheromone-mediated mass attacks which, in concert with its vectored fungi, may kill trees over large areas (Aukema *et al.*, 2008). On-going research is aimed at elucidating the roles that various factors play in MPB outbreaks, such as predators, pathogens, heterogeneity of habitat, climate, reproduction and dispersal. Identifying and understanding the key factors

Table 2. Average number of correctly identified zero and non-zero regression coefficients $\{\beta_j\}$ and auto-regressive coefficients $\{\theta_k\}$ by using the proposed LARS_m method with either one or two regularization parameters and compared with three alternatives: LARS_1 with either one or two regularization parameters, the LQA and selection of regression coefficients only†

Method	*n*	*Number of non-zero* β_j	*Number of zero* β_j	*Number of non-zero* θ_k	*Number of zero* θ_k
LARS_m,	25	3.98	1.26	0.82	1.42
2-regular	100	4.00	2.52	1.00	3.51
	225	4.00	2.65	1.00	3.65
LARS_m,	25	3.99	1.11	0.98	0.48
1-regular	100	4.00	2.48	1.00	1.21
	225	4.00	2.53	1.00	1.30
LARS_1,	25	3.98	1.04	0.83	1.31
2-regular	100	4.00	1.78	1.00	2.80
	225	4.00	1.83	1.00	3.25
LARS_1,	25	3.99	0.92	0.98	0.37
1-regular	100	4.00	1.80	1.00	0.73
	225	4.00	1.69	1.00	0.78
LQA	25	3.98	1.41	0.80	1.75
	100	4.00	2.55	1.00	3.10
	225	4.00	2.57	1.00	3.51
LARS_m,	25	3.96	1.84	0.81	2.50
regression only	100	4.00	2.64	1.00	3.84
	225	4.00	2.60	1.00	3.85
LARS_1,	25	3.98	1.32	0.80	2.49
regression only	100	4.00	1.78	1.00	3.81
	225	4.00	1.81	1.00	3.86
Truth		4	3	1	4

†The sample sizes are $n = 25, 100, 225$ and the model is SAR.

could ultimately result in reliable predictive models that would greatly facilitate management and planning of pine forests.

The data example that is used for illustration here is a subset of the MPB data on a 10×10 grid of cells averaging 12 km $\times$ 12 km in size overlaying the Chilcotin plateau (Aukema *et al.*, 2008). The response variable is the intensity of MPB infestation within a cell. The covariates comprise topographical and climatic variables. In particular, the topography is based on a digital elevation map. It is plausible that higher elevations are associated with fewer outbreaks of MPB since these regions are associated with a cooler climate and less pine. The climatic variables are temperature in the previous calendar year (minimum, maximum and mean), mean August temperature, accumulated degree days above 5.5 °C from the previous August to the current July (DD) or from the previous August to the end of the growing season (DDEG) and the amount of precipitation, all of which are defined according to known biology and phenology of the insect. For example, warmer temperatures are associated with more outbreaks of MPB, since adult insects complete their development and fly in search of new trees in response to summer temperature thresholds.

Furthermore, it is important to account for any remaining spatial dependence in the regression. This would not only ensure that statistical inference of the regression coefficients is valid but would also be of scientific interest to quantify such spatial dependence. A highly plausible explanation for spatial dependence is dispersal of the MPB populations across space as,

Table 3. Non-zero APMLE and standard deviation (SD) of model parameters in SAR models A or B for the MPB data example

Variable	*Results for model A*			*Results for model B*		
	Parameter	*APMLE*	*SD*	*Parameter*	*APMLE*	*SD*
Covariates						
Elevation	β_1	—	—	β_1	—	—
Temperature minimum	β_2	—	—	β_2	—	—
Temperature maximum	β_3	−6.40	14.45	β_3	−6.30	13.50
Temperature mean	β_4	6.64	11.34	β_4	6.44	10.65
August temperature mean	β_5	—	—	β_5	—	—
DD	β_6	—	—	β_6	—	—
DDEG	β_7	—	—	β_7	—	—
Precipitation	β_8	1.63	0.78	β_8	2.68	0.85
Order of neighbourhoods						
1st	θ_1	0.15	0.03	$\theta_{1,\updownarrow}$	0.18	0.04
				$\theta_{1,\leftrightarrow}$	0.12	0.05
2nd	θ_2	—	—	θ_2	—	—
3rd	θ_3	0.08	0.04	$\theta_{3,\updownarrow}$	0.13	0.05
				$\theta_{3,\leftrightarrow}$	—	—
4th	θ_4	—	—	θ_4	—	—
5th	θ_5	—	—	θ_5	—	—
Variance	σ^2	8.62	1.28	σ^2	8.24	1.22
BIC		352.45			354.44	

during outbreaks of MPB, beetle populations are capable of dispersal at large scale. For these purposes, we consider neighbourhood structures from the first to the fifth order. Both SAR and CAR models are fitted as the covariance for the error term. On the basis of BIC-values, however, it appears that the SAR model performs consistently better than the CAR model. Thus the discussion henceforth will be focused on the SAR model.

A spatial adaptive lasso using the LARS_m algorithm is applied to select both the covariates and the neighbourhood orders. The results are shown as model A in Table 3. Three covariates are selected, namely maximum temperature, mean temperature and precipitation. There is a positive relationship between intensity of infestation and precipitation and no significant relationship with the other covariates including elevation. For spatial auto-correlation, the first- and third-order neighbourhoods are selected, but not the second-, fourth- and fifth-order neighbourhoods.

We further evaluate anisotropy by partitioning the first- and third-order neighbourhood into $\mathcal{N}_k(i) = \mathcal{N}_{k,\updownarrow}(i) \cup \mathcal{N}_{k,\leftrightarrow}(i)$ for $k = 1, 3$, i.e. we consider different auto-regressive coefficients for the north–south direction *versus* the west–east direction. The results are shown as model B in Table 3. The same three covariates are selected whereas the second-, fourth- and fifth-order neighbourhoods are not selected, as before. Neighbourhoods in both directions are selected on the first order, but only the north–south direction on the third order. Since the BIC-value for this model is slightly higher, model A seems to be more adequate than model B. We present the results here nonetheless, in part to illustrate the flexibility in specifying the neighbourhood structures.

An additional table is given in the Web-based supplementary materials that shows parameter estimation using LARS_1 and LQA. The results are generally similar, although the BIC-values become larger, indicating somewhat worse performance.

7. Discussion

In this paper, we have considered spatial linear models for lattice data, which have two additive components of a linear regression and an error term. For the covariance of the error term, we have focused on CAR and SAR models. For geostatistical data, the approach that we have proposed may still be applicable, via discretization of the continuous spatial domain. For example, Zhu and Liu (2009) used an SAR-type lattice model to analyse a geostatistical rainfall data set. Moreover, our methodology may be extended to deal with other types of dependence. For example, it would be interesting to consider further the non-stationary covariance of the error term. Zhu and Liu (2009) allowed non-stationarity via a non-parametric approach but required replications of the spatial data for estimation. It would be interesting to develop general classes of non-stationary models that would not require such replications. Finally, it would also be interesting to investigate dependence structure that is not necessarily spatial. We leave these for future investigation.

Acknowledgements

The authors are grateful to the Joint Editor, an Associate Editor and two referees for their helpful comments and suggestions, as well as Dr Brian Aukema for providing the MPB data set for illustration.

References

Aukema, B. H., Carroll, A. L., Zheng, Y., Zhu, J., Raffa, K. F., Moore, R. D. and Stahl, K. (2008) Movement of outbreak populations of mountain pine beetle: influences of spatiotemporal patterns and climate. *Ecography*, **31**, 348–358.

Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. R. Statist. Soc.* B, **36**, 192–236.

Cressie, N. (1993) *Statistics for Spatial Data*, revised edn. New York: Wiley.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least angle regression (with discussion). *Ann. Statist.*, **32**, 407–499.

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Ass.*, **96**, 1348–1360.

Hesterberg, T., Choi, N., Meier, L. and Fraley, C. (2008) Least angle and L_1 penalized regression: a review. *Statist. Surv.*, **2**, 61–93.

Huang, H.-C., Hsu, N.-J., Theobald, D. and Breidt, F. J. (2010) Spatial LASSO with applications to GIS model selection. *J. Computnl Graph. Statist.*, to be published.

Schabenberger, O. and Gotway, C. A. (2005) *Statistical Methods for Spatial Data Analysis*. Boca Raton: Chapman and Hall.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **58**, 267–288.

Wang, H., Li, G. and Tsai, C.-L. (2007a) Regression coefficients and autoregressive order shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **69**, 63–78.

Wang, H., Li, R. and Tsai, C.-L. (2007b) Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, **94**, 553–568.

Zhu, Z. and Liu, Y. (2009) Estimating spatial covariance using penalized likelihood with weighted L_1 penalty. *J. Nonparam. Statist.*, **21**, 925–942.

Zou, H. (2006) The adaptive LASSO and its oracle properties. *J. Am. Statist. Ass.*, **101**, 1418–1429.

Zou, H. and Li, R. (2008) One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *Ann. Statist.*, **36**, 1509–1566.