Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2017)
79, *Part* 3, *pp.* 779–800

Additive model building for spatial regression

Siddhartha Nandy,

Michigan State University, East Lansing, USA

Chae Young Lim

Seoul National University, Korea

and Tapabrata Maiti

Michigan State University, East Lansing, USA

[Received June 2014. Final revision May 2016]

Summary. Spatial regression is an important predictive tool in many scientific applications and an additive model provides a flexible regression relationship between predictors and a response variable. We develop a regularized variable selection technique for building a spatial additive model. We find that the methods developed for independent data do not work well for spatially dependent data. This motivates us to propose a spatially weighted l_2-error norm with a group lasso type of penalty to select additive components in spatial additive models. We establish the selection consistency of the approach proposed where the penalty parameter depends on several factors, such as the order of approximation of additive components, characteristics of the spatial weight and spatial dependence. An extensive simulation study provides a vivid picture of the effects of dependent data structure and choice of a spatial weight on selection results as well as the asymptotic behaviour of the estimators. As an illustrative example, the method is applied to lung cancer mortality data over the period of 2000–2005, obtained from the 'Surveillance, epidemiology, and end results' programme, National Cancer Institute, USA.

Keywords: Additive models; Group lasso; High dimension; Spatial dependence; Spatial regression; Variable selection

1. Introduction

It is important yet fairly challenging to identify important factors that explain certain phenomena such as climate change, economic volatility, ecological dynamics and disease progress. In such applications, spatially dependent data are often observed and spatial regression is a natural tool for data analysis. An additive model provides a flexible regression relationship and has been proven to be effective for regression-based prediction. The models that have been developed for independent data are statistically inefficient for dealing with spatial data. Thus, the statistical models accounting for spatial dependence have received considerable attention over the last few decades.

A common feature for spatial data is spatial dependence between sampling sites. Generally, we assume that the dependence between two data points at two sampling sites decreases as the distance between two sites increases. At each sampling location, we observe a quantity of interest (the response variable) and additional information (covariates or predictors) that could

Address for correspondence: Chae Young Lim, Department of Statistics, Seoul National University, Gawnak-ro 1, Gawnak-gu, Seoul, Korea 08826.
E-mail: twinwood@snu.ac.kr

1369–7412/17/79779

affect the response. A natural approach to identify contributing factors that influence a response variable is selection of covariates in a regression set-up, which is commonly known as variable selection. Among many variable selection techniques, regularization techniques (e.g. Tibshirani (1996)) have received huge attention in recent years.

The current literature on variable selection concentrates heavily on regression models that are generally appropriate for independent observations. The methods without any adaptation are not expected to work well for spatially dependent data. Further, theoretical justification for spatial data requires special attention, which is also true for variable selection. There are studies on variable selection for time series data. Typical time series data can be viewed as a special case of spatial lattice data by reducing the dimension of an observation domain to 1 and assuming that the data are observed evenly over time. Wang *et al.* (2007) studied selection of regression coefficients and auto-regressive order via the lasso for regression models with auto-regressive errors. Nardi and Rinaldo (2011) considered the lasso for auto-regressive process modelling. Hsu *et al.* (2008) applied the lasso to select the subset for vector auto-regressive processes. Xu *et al.* (2012) studied variable selection for auto-regressive models with infinite variance. Whereas these approaches deal with a specific dependence structure (auto-regressive structure), Gupta (2012) investigated variable selection for weakly dependent time series data under a linear regression model.

Variable selection or model selection for spatial data is relatively new. Hoeting *et al.* (2006) derived Akaike's information criterion for a geostatistical model and used it for selecting explanatory variables under spatial correlation. Huang and Chen (2007) introduced a model selection criterion with generalized degrees of freedom for selecting a spatial prediction model. Huang *et al.* (2010b) considered a spatial lasso for selecting covariates and spatial neighbourhoods with a known spatial dependence structure but no theoretical investigation was made. Wang and Zhu (2009) considered penalized least squares for geostatistical data with various penalty functions and investigated their theoretical properties.

In likelihood-based approaches, Zhu *et al.* (2010) considered selection of spatial linear models together with a spatial neighbourhood structure for spatial lattice data by using an adaptive lasso penalty. Chu *et al.* (2011) investigated variable selection for spatial linear models for geostatistical data by using a penalized maximum likelihood method. They considered an approximate penalized maximum likelihood approach with a tapered spatial covariance function. Reyes *et al.* (2012) extended the approach of Zhu *et al.* (2010) for spatial–temporal lattice models. For spatial binary data, Fu *et al.* (2013) considered selection in autologistic regression models by using a penalized pseudolikelihood.

In this paper, we consider an additive regression model with spatially dependent error. Consider $\{Y(\mathbf{s}); \mathbf{s} \in \mathbb{R}^d\}$ to be a spatial process on $\mathbb{R}^d$ and $\{\mathbf{X}(\mathbf{s}) = (X_1(\mathbf{s}), \ldots, X_J(\mathbf{s})); \mathbf{s} \in \mathbb{R}^d\}$ to be a J-dimensional vector which can be stochastic, if necessary. We consider a spatial additive model given in model (1) in Section 2 with overall mean μ and f_j being unknown functions describing the relationship between Y and X_j. We assume that the error process is a mean 0 stationary Gaussian random field with covariance function $\delta(\mathbf{h})$. J could be larger than the sample size. Our objective is to select the 'effective' f_js.

For selection and estimation of non-linear components f_j, Huang *et al.* (2010a) proposed the adaptive group lasso (AGL) for the additive model (1) *but* with independent errors. Their work is based on a spline approximation of non-linear components which led to rewriting the mean of model (1) as a linear regression model with spline coefficients. Hence the problem of selecting a component in an additive model is transformed into selecting a group of variables in a linear regression with predefined group memberships. Meier *et al.* (2009) also considered variable selection for high dimensional additive models with a sparsity–smoothness penalty, which controls both sparsity as well as smoothness in spline approximations. Several other works on the selection of additive models or additive non-parametric regression comprise Antoniadis and Fan (2001), Lin and Zhang (2006), Ravikumar *et al.* (2009) and Lai *et al.* (2012), etc. These works assumed independent error distributions. Kneib *et al.* (2009) considered variable selection by using a penalized spline approach for geoadditive models without theoretical justification. To the best of our knowledge, there is no work on spatial additive model selection with a theoretical justification.

First, we empirically examined a group lasso (GL) approach developed for independent data to select non-zero components in an additive model when the errors are spatially dependent. For the additive model (1), we considered $J=10$ with two true non-zero components: $f_1(x)=\sin(x)$ and $f_2(x)=x$. We considered $m\times m$ unit square lattices with $m=6, 12, 24$. For the spatial errors, we used a Gaussian distribution with an exponential covariance function: $\delta(\mathbf{h})=\exp(-\rho|\mathbf{h}|)$ and $\rho=0.5$. We generated 400 data sets. Since the spatial dependence can also be captured by a function of the location in the mean, we also investigated two different intercepts in addition to the additive component model (1): one is a constant and the other is a non-parametric function of the location by using a spline approximation. For a non-parametric function of location $\mathbf{s}=(s_1, s_2)$, we considered an additive structure in terms of s_1 and s_2. Table 1 shows the average and standard deviation of the number of selected components when the independence approach is used. The regularization parameter (or penalty parameter) is chosen as recommended by Huang *et al.* (2010a). The number of covariates selected is much larger than the true number of non-zero components for both cases at various sample sizes. Although a function of location as a mean component helps to improve the selection results in this simulation study, it is not satisfactory. The performance did not improve even with a larger sample size. This leads us to investigate a variable selection method that is suitable for spatial additive models. A comprehensive simulation study is given in Section 4.

For spatial additive models, we maintain the idea of approximating f_j by spline functions and a GL penalty for sparsity. Then, we introduce a spatial weight matrix in the objective function. The choice of a weight matrix is motivated by the concept of weighted least squares. The dependence in the random error is compensated for by a spatial weight matrix. We develop asymptotic theory for selection consistency of non-zero components in the Gaussian spatial additive model. The spatial dependence structures that we consider are common in modelling spatial data and are valid for a wide range of applications (e.g. Cressie (2015) and Stein (1999)). Variable selection is sensitive to the choice of penalty parameter. We found that the theoretical

Table 1. Average and standard deviation for the number of selected covariates by using 400 data sets from the covariance function $\delta(\mathbf{h})=\exp(-\rho|\mathbf{h}|)$ with $\rho=0.5$†

m	*Constant*		*Function of the location*	
	Average	*Standard deviation*	*Average*	*Standard deviation*
6	5.64	1.74	3.61	1.23
12	6.61	1.60	4.31	1.62
24	7.22	1.38	5.91	2.01

†The true number of non-zero components is 2. $m\times m$ unit square lattices are considered.

lower bound for the penalty parameter depends on spatial information of the data as well as a spatial weight matrix.

Our asymptotic results allow an identity matrix as a choice of the spatial weight matrices, although this may not be the optimal choice. This choice does not make our method identical to independent data analysis methods, such as Huang *et al.* (2010a), since the upper bounds of the rate of convergence and the lower bound of the penalty parameter are different.

The rest of the paper is organized as follows. Section 2 describes the proposed approach for selecting and estimating non-zero components in an additive model with spatially dependent errors. Section 3 discusses the main theoretical results for asymptotic properties of our proposed estimators. Section 4 contains simulation results along with a real data example for illustration. Finally, we make some concluding remarks in Section 5. Proofs of all the theorems are provided in Appendix A. Proofs of related lemmas, an extension of the theoretical results and additional simulation results are given in on-line supplementary material. All numerical study was performed by code written in statistical software R. Example code is available from

```
http://wileyonlinelibrary.com/journal/rss-datasets
```

2. Method for selecting components in spatial additive models

We consider the spatial additive model

$$Y(\mathbf{s})=\mu+\sum_{j=1}^{J} f_j\{X_j(\mathbf{s})\}+\epsilon(\mathbf{s}), \qquad \forall\, \mathbf{s}\in\mathbb{R}^d, \tag{1}$$

where μ is the overall mean, f_j is an unknown function describing the relationship between Y and X_j, and $\{\epsilon(\mathbf{s});\mathbf{s}\in\mathbb{R}^d\}$ is a zero-mean stationary Gaussian random field with a covariance function $\delta(\mathbf{h})$. Suppose that $(Y(\mathbf{s}),\mathbf{X}(\mathbf{s}))$ are observed at n different locations lying in sampling region $\mathbf{D}_n\subset\mathbb{R}^d$. Let $\mathbf{S}$ be the set of sampling locations. We use a spline approximation for f_j in the additive models:

$$f_j\{X_j(\mathbf{s})\}\approx f_{nj}\{X_j(\mathbf{s})\}:=\sum_{l=1}^{m_n}\beta_{jl}\mathbb{B}_l\{X_j(\mathbf{s})\}, \qquad \text{for } j=1,\ldots,J, \tag{2}$$

where the $\mathbb{B}_l(\cdot)$s are normalized B-spline bases, and the β_{jl}s are called control points (Schumaker, 2007). The approximation (2) is based on theory which states that every smooth function can be uniquely represented by a linear combination of B-splines. Then, model (1) is approximated as

$$Y(\mathbf{s})=\mu+\sum_{j=1}^{J}\sum_{l=1}^{m_n}\beta_{jl}\mathbb{B}_l\{X_j(\mathbf{s})\}+\pi(\mathbf{s}), \qquad \text{for } \mathbf{s}\in\mathbf{S}, \tag{3}$$

where $\pi(\mathbf{s})=\epsilon(\mathbf{s})+\theta(\mathbf{s})$ with $\theta(\mathbf{s})=\Sigma_{j=1}^{J}[f_j\{X_j(\mathbf{s})\}-f_{nj}\{X_j(\mathbf{s})\}]$ and also define $\theta=(\theta(\mathbf{s});\mathbf{s}\in\mathbf{S})'$. Model (3) can be written in matrix form $\mathbf{Y}=\mu+\mathbb{B}\boldsymbol{\beta}+\pi$, where $\mathbf{Y}=(Y(\mathbf{s}),\mathbf{s}\in\mathbf{S})'$, $\boldsymbol{\beta}=(\boldsymbol{\beta}'_1,\boldsymbol{\beta}'_2,\ldots,\boldsymbol{\beta}'_J)'$, $\mathbb{B}$ is the design matrix constructed by spline functions and $\boldsymbol{\pi}=(\pi(\mathbf{s}),\mathbf{s}\in\mathbf{S})'$.

For $\mathbf{v}=(\mathbf{v}'_1,\mathbf{v}'_2,\ldots,\mathbf{v}'_J)'$, where the $\mathbf{v}_j$s are vectors and $\boldsymbol{\psi}=(\psi_1,\psi_2,\ldots,\psi_J)'$, we define a weighted l_1-l_2-norm, $\|\mathbf{v}\|_{2,1,\boldsymbol{\psi}}=\Sigma_{j=1}^{J}\psi_j\|\mathbf{v}_j\|_2$, where $\|\cdot\|_2$ is the l_2-norm of a vector. This is a weighted l_1-norm of $(\|\mathbf{v}_1\|_2,\ldots,\|\mathbf{v}_J\|_2)$ with a weight vector $\boldsymbol{\psi}$. Then, we propose the following weighted l_1-l_2-penalized least squares objective function, weighted by spatial weight matrix $\boldsymbol{\Sigma}_{\mathrm{W}}$:

$$\mathbf{Q}_n(\boldsymbol{\beta},\lambda_n)=(\mathbf{Y}-\mu-\mathbb{B}\boldsymbol{\beta})'\boldsymbol{\Sigma}_W^{-1}(\mathbf{Y}-\boldsymbol{\mu}-\mathbb{B}\boldsymbol{\beta})+\lambda_n\|\boldsymbol{\beta}\|_{2,1,\boldsymbol{\psi}_n}, \tag{4}$$

where λ_n is a regularization parameter, $\boldsymbol{\psi}_n = (\psi_{n1}, \psi_{n2}, \ldots, \psi_{nJ})'$ is a suitable choice of a weight vector for the penalty term and $\boldsymbol{\beta}$ is the $(Jm_n \times 1)$-dimensional vector of control points introduced in approximation (2). $\boldsymbol{\Sigma}_{\mathrm{W}}$ is a known positive definite spatial weight matrix where more weights are given if two locations are closer and vice versa. A discussion on spatial weight matrices is provided in Section 3.1.

We allow the possibility that the regularization parameter λ_n and the weight vector $\boldsymbol{\psi}_n$ can depend on the sample size n. To avoid any identifiability issue, we assume that $\mathbb{E}\{f_j(X_j)\} = 0, \forall\ 1 \leqslant j \leqslant J$, which leads us to assume that

$$\sum_{l=1}^{m_n} \beta_{jl}\mathbb{B}_l\{X_j(\mathbf{s})\} = 0, \qquad \forall\ 1 \leqslant j \leqslant J, \quad \mathbf{s} \in \mathbf{S}. \tag{5}$$

Combining expressions (4) and (5), we have the unconstrained objective function given by

$$\mathbf{Q}_n(\boldsymbol{\beta}, \lambda_n) = (\mathbf{Y}^{\mathrm{c}} - \mathbb{B}^{\mathrm{c}}\boldsymbol{\beta})'\boldsymbol{\Sigma}_{\mathrm{W}}^{-1}(\mathbf{Y}^{\mathrm{c}} - \mathbb{B}^{\mathrm{c}}\boldsymbol{\beta}) + \lambda_n\|\boldsymbol{\beta}\|_{2,1,\boldsymbol{\psi}_n}, \tag{6}$$

where $\mathbf{Y}^{\mathrm{c}} = (Y^{\mathrm{c}}(\mathbf{s}), \mathbf{s} \in \mathbf{S})' = (Y(\mathbf{s}) - \bar{Y}, \mathbf{s} \in \mathbf{S})'$ with $\bar{Y} = (1/n)\,\Sigma_{\mathbf{s}\in\mathbf{S}}Y(\mathbf{s})$ and $\mathbb{B}^{\mathrm{c}} = (\mathbb{B}^{\mathrm{c}}_1, \mathbb{B}^{\mathrm{c}}_2, \ldots, \mathbb{B}^{\mathrm{c}}_J)$ is the design matrix with an $n \times m_n$ matrix $\mathbb{B}^{\mathrm{c}}_j$. Each row of $\mathbb{B}^{\mathrm{c}}_j$ is $(\mathbb{B}^{\mathrm{c}}_1\{X_j(\mathbf{s})\}, \ldots, \mathbb{B}^{\mathrm{c}}_{m_n}\{X_j(\mathbf{s})\})$ with $\mathbb{B}^{\mathrm{c}}_l\{X_j(\mathbf{s})\} = \mathbb{B}_l\{X_j(\mathbf{s})\} - (1/n)\Sigma_{\mathbf{s}'\in\mathbf{S}}\mathbb{B}_l\{X_j(\mathbf{s}')\}$.

First, we obtain an estimate of $\boldsymbol{\beta}$ by minimizing the objective function $\mathbf{Q}_{n1}(\boldsymbol{\beta}, \lambda_{n1}) := \mathbf{Q}_n(\boldsymbol{\beta}, \lambda_{n1})$ with $\psi_{nj} = 1$, for all $j = 1, \ldots, J$. We call this estimate the GL estimate $\hat{\boldsymbol{\beta}}_{\mathrm{GL}}(\lambda_{n1})$, and the corresponding objective function the GL objective function. Note that $\|\boldsymbol{\beta}\|_{2,1,\mathbf{1}} = \Sigma_{j=1}^{J}\|\boldsymbol{\beta}_j\|_2$. To improve the selection, we use the following updated weights from $\hat{\boldsymbol{\beta}}_{\mathrm{GL}}$,

$$\psi_{nj} = \begin{cases} 1/\|\hat{\boldsymbol{\beta}}_{\mathrm{GL},j}\|_2, & \text{if } \|\hat{\boldsymbol{\beta}}_{\mathrm{GL},j}\|_2 > 0, \\ \infty, & \text{if } \|\hat{\boldsymbol{\beta}}_{\mathrm{GL},j}\|_2 = 0, \end{cases} \tag{7}$$

in an objective function which is called an AGL objective function, i.e. we define an AGL objective function $\mathbf{Q}_{n2}(\boldsymbol{\beta}, \lambda_{n2}) := \mathbf{Q}_n(\boldsymbol{\beta}, \lambda_{n2})$ with $\boldsymbol{\psi}_n$ given in expression (7). The estimate from this updated objective function, $\hat{\boldsymbol{\beta}}_{\mathrm{AGL}}(\lambda_{n2}) = \arg\min_{\boldsymbol{\beta}}\mathbf{Q}_{n2}(\boldsymbol{\beta}, \lambda_{n2})$, as a function of λ_{n2} is referred to as an AGL estimate.

We define $\infty \times 0 = 0$, so components that are not selected by the GL method are not included for the AGL method. Also, because of the nature of weights in the penalty term, the AGL objective function puts a higher penalty on components with smaller l_2-norm and lower penalty on components with larger l_2-norm of the GL estimates. Hence, the components with larger GL estimates have a higher chance of being selected in the final model. Finally, the AGL estimates for μ and f_j are given by $\hat{\mu} = \bar{Y} = (1/n)\Sigma_{\mathbf{s}\in\mathbf{S}}Y(\mathbf{s})\,\hat{f}_{\mathrm{AGL},j}\{X_j(\mathbf{s})\} = \Sigma_{l=1}^{m_n}\hat{\beta}_{\mathrm{AGL},jl}\mathbb{B}_l\{X_j(\mathbf{s})\}$ for all $j = 1, \ldots, J$.

The computation of $\boldsymbol{\Sigma}_{\mathrm{W}}^{-1}$ in equation (6) may cause computational complication particularly for *large n*. To avoid such a situation, we reformulate equation (6) by using a Cholesky decomposition of $\boldsymbol{\Sigma}_{\mathrm{W}}^{-1}$. Let $\boldsymbol{\Sigma}_{\mathrm{W}}^{-1} = \mathbf{L}\mathbf{L}'$, where $\mathbf{L}$ is a lower triangular matrix. Then, equation (6) can be rewritten as $\mathbf{Q}_n(\boldsymbol{\beta}, \lambda_n) = (\mathbf{Z}^{\mathrm{c}} - \mathbb{D}^{\mathrm{c}}\boldsymbol{\beta})'(\mathbf{Z}^{\mathrm{c}} - \mathbb{D}^{\mathrm{c}}\boldsymbol{\beta}) + \lambda_n\|\boldsymbol{\beta}\|_{2,1,\boldsymbol{\psi}_n}$, where $\mathbf{Z}^{\mathrm{c}} = \mathbf{L}'\mathbf{Y}^{\mathrm{c}}$ and $\mathbb{D}^{\mathrm{c}} = \mathbf{L}'\mathbb{B}^{\mathrm{c}}$ with $\mathbb{D}^{\mathrm{c}}_j = \mathbf{L}'\mathbb{B}^{\mathrm{c}}_j$. Then, the objective function becomes the function with no spatial weight matrix with a new response variable $\mathbf{Z}$ so that we can adopt an available algorithm for a GL method with independent errors. Because of the known spatial weight matrix the Cholesky decomposition is required to be done only once.

3. Main theoretical results

In this section, we present asymptotic properties of the GL and AGL estimators that were introduced in Section 2. We start by introducing some notation. Let A_0 and A_* be the sets of zero and non-zero components respectively. Without loss of generality, we consider $A_* = \{1, 2, \ldots, q\}$ and $A_0 = \{q+1, \ldots, J\}$. We assume that there exists $\tilde{A}_0$ that satisfies $\Sigma_{j \in \tilde{A}_0} \|\boldsymbol{\beta}_j\|_2 \leqslant \eta_1$ for some $\eta_1 \geqslant 0$ and let $\tilde{A}_* = \{1, 2, \ldots, J\} \setminus \tilde{A}_0$. Existence of $\tilde{A}_0$ is referred to as the generalized sparsity condition (Zhang and Huang, 2008). First, we consider the selection and estimation properties of the GL estimator. Let $\tilde{A}_\beta$ be the index set of non-zero GL estimates for $\boldsymbol{\beta}_j$ and $\hat{A}_f$ be the index set of non-zero GL estimators for f_j. Necessary assumptions to study asymptotic properties are given in assumption 1.

Assumption 1.

(a) Among J covariates, the number of non-zero components, q, is fixed and there is a constant $k_f > 0$ such that $\min_{1 \leqslant j \leqslant q} \|f_j\|_2 \geqslant k_f$.
(b) There exists $v > 0$ such that $\min_{\mathbf{s} \neq \mathbf{s}' \in \mathbf{S}} \|\mathbf{s} - \mathbf{s}'\|_2 > v$, where $\|\cdot\|_2$ is the l_2-norm for a vector.
(c) The random vector $\boldsymbol{\epsilon} = \{\epsilon(\mathbf{s}), \mathbf{s} \in \mathbf{S}\} \sim \text{Gaussian}(0, \boldsymbol{\Sigma}_\text{T})$, where $\boldsymbol{\Sigma}_\text{T}$ is constructed by a stationary covariance function $\delta_\text{T}(\mathbf{h})$ which satisfies $\int_{\mathbf{D}_n} \delta_\text{T}(\mathbf{h})\, \mathrm{d}\mathbf{h} < \infty$. $\mathbf{D}_n \subset \mathbb{R}^d$ is the sampling region that contains the sampling locations $\mathbf{S}$. We assume that the origin of $\mathbb{R}^d$ is in the interior of $\mathbf{D}_n$ and $\mathbf{D}_n$ is increasing with n.
(d) $f_j \in \mathcal{F}$ and $\mathbb{E}\{f_j(X_j)\} = 0$ for $j = 1, \ldots, J$, where $\mathcal{F} = \{f \,|\, |f^{(k)}(s) - f^{(k)}(t)| \leqslant C|\, s - t|^{\nu}$, $\forall s, t \in [a, b]\}$ for some non-negative integer k and $\nu \in (0, 1]$. Also suppose that $\tau = k + \nu > 1$.
(e) The covariate vector $\mathbf{X}$ has a bounded continuous density function $g_j(x)$ of X_j on a bounded domain $[a, b]$ for $j = 1, \ldots, J$.
(f) $m_n = O(n^{\gamma})$ with $\frac{1}{6} \leqslant \gamma = 1/(2\tau + 1) < \frac{1}{3}$.
(g) $\boldsymbol{\Sigma}_\text{W}$ is constructed by a stationary covariance function $\delta(\mathbf{h})$ that satisfies the same condition as $\delta_\text{T}(\mathbf{h})$ in part (c) and $\kappa(\boldsymbol{\Sigma}_\text{W}) = \kappa(\boldsymbol{\Sigma}_\text{W}^{-1}) \leqslant M$ for some $M < \infty$, where κ is the condition number of a matrix.

Assumption 1, part (a), indicates that we need strong signals for non-zero components to distinguish them from the noise. Assumption 1, part (b), implies that we consider increasing domain asymptotics (Stein, 1999) for our large sample properties, which is a common sampling assumption for the asymptotic theory of spatial statistics. Assumption 1, part (c), specifies distributional assumptions of spatial error and its spatial dependence. Commonly available stationary spatial covariance models satisfy the integrability assumption. For example, popular spatial covariance functions such as exponential, Matérn and Gaussian covariance functions are all integrable. For explicit expression of such covariance functions, see the on-line supplementary material. We assume that $\mathbf{D}_n$ contains the origin to have spatial lag $\mathbf{h}$ and observation location $\mathbf{s}$ on the same spatial domain. The stationary spatial covariance function $\delta(\mathbf{h})$ gives a marginal variance at $\mathbf{h} = \mathbf{0}$ (i.e. $\mathbf{s} = \mathbf{s}'$) and decreases towards zero as $\|\mathbf{h}\| \to \infty$. The condition on $\delta(\mathbf{h})$ in assumption 1, part (c), becomes meaningful by assuming that $\mathbf{D}_n$ contains the origin. For example, the condition does not guarantee whether $\delta(\mathbf{h})$ is integrable if we do not assume that $\mathbf{D}_n$ contains the origin.

Assumption 1, part (d), considers that f_js are in the class of functions defined on $[a, b]$ such that the kth derivative satisfies the Lipschitz condition of order ν and the zero expectation condition is needed to avoid an identifiability issue. Assumption 1, part (e), is needed to have a spline approximation for additive components. Assumption 1, part (f), is related to the number of B-spline bases to approximate additive components. The parameter γ controls the smoothness of additive components, where imposing an upper and lower bound on γ implies that those functions can be neither too smooth nor too wiggly. If functions are too smooth, it would be difficult to detect those distinctly from the overall mean, whereas, if the functions are too wiggly, it would be difficult to detect those distinctly from random noise.

Assumption 1, part (g), implies that a spatial weight matrix $\boldsymbol{\Sigma}_\text{W}$ is a well-conditioned matrix.

Note that we consider a stationary spatial covariance function to construct a spatial weight matrix. Under assumption 1, part (b), we can see that the smallest eigenvalue of the spatial weight matrix constructed by several spatial covariance functions is bounded away from zero (see Wendland (2005)). In contrast, the largest eigenvalue of a spatial weight matrix is related to the norm of the matrix. By Geršgorin's theorem (Horn and Johnson, 1985), we can show that $\rho_{\max}(\Sigma_{\mathrm{W}}) \leqslant \max_j \Sigma_k |\Sigma_{\mathrm{W},jk}| = \max_j \Sigma_k \delta(\mathbf{s}_j - \mathbf{s}_k)$, where $\Sigma_{\mathrm{W},jk}$ is the (j,k)th entry of Σ_{W}. This is bounded by a finite constant which is independent of n for the spatial weight matrix from stationary integrable covariance functions.

Now, we introduce the consistency result for the GL estimator.

Theorem 1. Suppose that the conditions in assumption 1 hold and

$$\lambda_{n1} > C\,\rho_{\max}(\mathbf{L})\surd\{nm_n \log(Jm_n)\}$$

for a sufficiently large constant C. Then, we have

(a)

$$\sum_{j=1}^{J} \|\hat{\boldsymbol{\beta}}_{\mathrm{GL},j} - \boldsymbol{\beta}_j\|_2^2 = \mathbf{O}_p\left\{ \frac{\rho^2_{\max}(\mathbf{L}) m_n^3 \log(Jm_n)}{n} + \frac{m_n}{n} + \frac{1}{m_n^{2\tau-1}} + \frac{4m_n^2\lambda_{n1}^2}{n^2} \right\},$$

(b) if $m_n^2\lambda_{n1}^2/n^2 \to 0$ as $n \to \infty$, all the non-zero components $\boldsymbol{\beta}_j$, $1 \leqslant j \leqslant q$, are selected with probability converging to 1.

The spatial dependence of the data contributes to the theoretical lower bound of the penalty parameter and the convergence rate by an additional m_n-term in contrast with the results from independent data analysis. Recall that m_n is the number of spline basis functions for approximating the f_js. This additional m_n comes from the bound of the expected value for a function of spatially dependent error (see lemma 3 in Appendix A). The spatial weight matrix also contributes to the theoretical lower bound of the penalty parameter and the convergence rate via the maximum eigenvalue of $\mathbf{L}$. Recall that $\mathbf{L}$ is the Cholesky decomposition component of Σ_{W}^{-1}. This implies that spatial dependence of the data and a spatial weight matrix affect the rate of convergence and the selection of components via the choice of the penalty parameter. All these additional quantities make the lower bound of the penalty parameter for spatial additive models larger compared with the independent data case.

Although the approach to prove the asymptotic properties that were stated above for spatial additive models is similar to that for additive models with independent errors, details are different because of spatial dependence as well as a spatial weight matrix in the method proposed. Proofs of theorem 1 as well as subsequent theorems are given in Appendix A. The next theorem provides consistency in terms of the estimated non-linear components $\hat{f}_j$.

Theorem 2. Suppose that the conditions in assumption 1 hold and if $\lambda_{n1} > C\,\rho_{\max}(\mathbf{L}) \times \surd\{nm_n \log(Jm_n)\}$ for a sufficiently large constant C. Then,

(a)

$$\|\hat{f}_{\mathrm{GL},j} - f_j\|_2^2 = \mathbf{O}_p\left\{ \frac{\rho^2_{\max}(\mathbf{L}) m_n^2 \log(Jm_n)}{n} + \frac{1}{n} + \frac{1}{m_n^{2\tau}} + \frac{4m_n\lambda_{n1}^2}{n^2} \right\}$$

for $j \in \tilde{A}_{\beta} \cup A_*$, where $\tilde{A}_{\beta}$ is the index set of non-zero GL estimates for $\boldsymbol{\beta}_j$, and,

(b) if $m_n\lambda_{n1}^2/n^2 \to 0$ as $n \to \infty$, all the non-zero components f_j, $1 \leqslant j \leqslant q$, are selected with probability converging to 1.

By theorem 2 with $\lambda_{n1} = O[\rho_{\max}(\mathbf{L})\surd\{nm_n \log(Jm_n)\}]$ and $m_n = O(n^{\gamma})$, we have

(i) $\|\hat{f}_{\mathrm{GL},j}-f_j\|_2^2=\mathbf{O}_p\{\rho^2_{\max}(\mathbf{L})n^{2\gamma}\log(Jm_n)/n\}$, for $j\in\tilde{A}_{\beta}\cup A_*$, and,

(ii) if $\rho^2_{\max}(\mathbf{L})\log(J)/n^{1-2\gamma}\rightarrow 0$ as $n\rightarrow\infty$ then, with probability converging to 1, all the non-zero components $f_j, 1\leqslant j\leqslant q$, are selected. This also implies that the number of additive components, J, can grow upto $\exp[o\{\rho^2_{\min}(\mathbf{L})n^{1-2\gamma}\}]$. In particular, if we need second-order differentiability of f_j, then $\tau=2$ implies that $\gamma=\frac{1}{5}$ and in this case J can be as large as $\exp[o\{\rho^2_{\min}(\mathbf{L})n^{3/5}\}]$. Keeping $\mathbf{L}$ fixed, we can see that J increases exponentially in n.

Next, we state the additional assumptions for the asymptotic properties of the AGL estimator.

Assumption 2.

(a) The initial estimators $\hat{\boldsymbol{\beta}}_{\mathrm{GL},j}$ are r_n consistent, $r_n\max_{1\leqslant j\leqslant J}\|\hat{\boldsymbol{\beta}}_{\mathrm{GL},j}-\boldsymbol{\beta}_j\|_2=\mathbf{O}_p(1)$, as $r_n\rightarrow\infty$, and there is a constant $k_b>0$ such that $\mathbb{P}(\min_{j\in A_*}\|\hat{\boldsymbol{\beta}}_{\mathrm{GL},j}\|_2\geqslant k_b b_{n1})\rightarrow 1$, where $b_{n1}=\min_{j\in A_*}\|\boldsymbol{\beta}_j\|_2\asymp m_n^{1/2}$ where, for two positive sequences a_n and b_n, $a_n\asymp b_n$ if there exist a_1 and a_2 such that $0<a_1<a_n/b_n<a_2<\infty$.

(b) $\surd\{\rho^2_{\max}(\mathbf{L})nm_n\log(s_nm_n)\}/\lambda_{n2}r_n+n^2/(\lambda^2_{n2}r^2_nm_n)+\lambda_{n2}m_n/n=o(1)$, where $s_n=J-|A_{**}|$. A_{**} is the set of indices that correspond to the components in the additive model which are correctly selected by the AGL approach. Mathematical definition is given in the proof of theorem 3.

Assumption 2, part (a), ensures the availability of an r_n-consistent estimator under certain regularity conditions. Also we can see that, under assumptions 1, parts (a)–(g), a suitable choice of r_n is $\{\rho^2_{\max}(\mathbf{L})m^3_n\log(Jm_n)/n\}^{-1/2}$. Then, assumption 2, part (b), can be replaced by

(b′) $\lambda_{n1}\surd m_n/\lambda_{n2}+\lambda_{n2}m_n/n=o(1)$.

A detailed derivation is given in the on-line supplementary material.

For selection consistency of the AGL estimator, we introduce '$\hat{\boldsymbol{\beta}}_{\mathrm{AGL}}=_0\boldsymbol{\beta}$', which means that $\mathrm{sgn}_0(\|\hat{\boldsymbol{\beta}}_{\mathrm{AGL},j}\|_2)=\mathrm{sgn}_0(\|\boldsymbol{\beta}_j\|_2)$ for all j, where $\mathrm{sgn}_0(\|x\|_2)$ equals 1 if $\|x\|_2>0$ and 0 if $\|x\|_2=0$. Define $J_0=|A_*\cup\{j:\|\hat{\boldsymbol{\beta}}_{\mathrm{AGL},j}\|_2>0\}|$. Note that J_0 is bounded by a finite number with probability converging to 1 by theorem 1.

Theorem 3. Suppose that the conditions in assumptions 1 and 2 are satisfied. Then,

(a) $\mathbb{P}(\hat{\boldsymbol{\beta}}_{\mathrm{AGL}}=_0\boldsymbol{\beta})\rightarrow 1$ and

(b)

$$\sum_{j=1}^{q}\|\hat{\boldsymbol{\beta}}_{\mathrm{AGL},j}-\boldsymbol{\beta}_j\|_2^2=\mathbf{O}_p\left\{\frac{\rho^2_{\max}(\mathbf{L})m^3_n\log(J_0m_n)}{n}+\frac{m_n}{n}+\frac{1}{m_n^{2\tau-1}}+\frac{4m^2_n\lambda^2_{n2}}{n^2}\right\}.$$

By theorem 3, we can show that the proposed AGL estimator of $\boldsymbol{\beta}$ can separate true zero components for the spatial additive model. Similarly to theorem 1, we have additional quantities on the right-hand side of the expression in theorem 3, part (b), which are due to the spatial dependence in errors as well as use of a spatial weight matrix. Since J_0 is bounded by a finite number with probability converging to 1, the convergence rate of the AGL estimator of $\boldsymbol{\beta}$ is faster than that of the GL estimator of $\boldsymbol{\beta}$. The next theorem shows that estimated components for f_js in the spatial additive model using the AGL estimator for $\boldsymbol{\beta}$ can identify non-zero components consistently. Also, the theorem provides the rate of convergence for the estimated components.

Theorem 4. Suppose that the conditions in assumptions 1 and 2 are satisfied. Then,

(a) $\mathbb{P}(\|\hat{f}_{\mathrm{AGL},j}\|_2>0, j\in A_*$ *and* $\|\hat{f}_{\mathrm{AGL},j}\|_2=0, j\notin A_*)\rightarrow 1$ and

(b)

$$\sum_{j=1}^{q} \| \hat{f}_{\mathrm{AGL},j} - f_j \|_2^2 = \mathbf{O}_p \left\{ \frac{\rho_{\max}^2(\mathbf{L}) m_n^2 \log(J_0 m_n)}{n} + \frac{1}{n} + \frac{1}{m_n^{2\tau}} + \frac{4 m_n \lambda_{n2}^2}{n^2} \right\}.$$

The upper bounds of the rates of convergence in theorems 1–4 show that the rates are slower than those for the independent data case. This is not surprising given that we are dealing with dependent data. Also, our theoretical results show that we need an improved lower bound of the penalty parameter for spatial additive models, which is critical since the penalty parameter is sensitive to the selection results in practice. This is supported by the simulation study in the next section where we can clearly see worse performance when we blindly apply the approach that was developed for independent data to select non-zero components.

We assumed that each additive component shares the same smoothness by part (d) in assumption 1. The results can be extended to allow different levels of smoothness for additive components without much difficulty. Details of such an extension are provided in the on-line supplementary material.

3.1. Selection of penalty parameter and spatial weight matrix

The selection result is sensitive to the choice of penalty parameter (or regularization parameter). In addition to the penalty parameter, the proposed approach for spatial additive models requires us to choose a spatial weight matrix as well. Theoretical results provide only the lower bound of the penalty parameter which involves the information for a spatial weight matrix through $\rho_{\max}(\mathbf{L})$. The approach to find an optimal penalty parameter in the penalized methods for independent data cannot be applied directly to our setting because of the spatial weight matrix. A complete theoretical investigation for finding an optimal value is interesting; however, it is beyond the scope of this paper. Also, a theoretically obtained optimal choice of the penalty parameter is not feasible in practice since it is only valid asymptotically and it often depends on unknown nuisance parameters in the true model (Fan and Tang, 2013). Thus, we demonstrate a practical way of selecting a penalty parameter guided by the theoretical results.

We assume that the spatial weight matrix is constructed by a class of stationary spatial covariance functions controlled by a parameter ρ, for simplicity. We call ρ the spatial weight parameter. Example classes of spatial covariance functions that satisfy the assumption 1, part (g), to construct a spatial weight matrix are the Gaussian covariance function and the inverse multiquadratic function, which are given in the on-line supplementary material. The selection problem is then reduced to selecting a spatial weight parameter. To choose a penalty parameter together, we consider a theoretical lower bound for the penalty parameter as the value of the penalty parameter at a given value of ρ so that one parameter (the spatial weight parameter) controls both the regularization level and spatial weight. Then, we adopt the generalized information criterion GIC (Nishii, 1984; Fan and Tang, 2013) as a measure to choose a spatial weight parameter, i.e. we find the ρ that minimizes

$$\mathrm{GIC}\{\lambda_n(\rho)\} = \log(\mathrm{RSS}) + \mathrm{df}_{\lambda_n} \frac{\log\{\log(n)\} \log(J)}{n}.$$

In practice, we consider a sequence of ρ and choose the one that minimizes GIC. While experimenting with different information criteria and comparing them with the existing cross-validation criterion that was suggested by Yuan and Lin (2006), we noted that we cannot have an initial least square estimator when $J > n$. Thus, we define the degrees of freedom df_{λ_n} as the total number of estimated non-zero components, i.e. $\mathrm{df}_{\lambda_n} = \hat{q}_{\lambda_n} m_n$ where $\hat{q}_{\lambda_n}$ is the active set of selected variables. For the independent data case, Huang *et al.* (2010a) suggested the extended Bayesian information criterion EBIC to choose a penalty parameter, which requires us to choose an additional parameter (ν in Huang *et al.* (2010a)). We found that a smaller value compared with the suggested value for the additional parameter works better in our setting from the simulation study. Given the sensitivity of EBIC with this additional parameter, we instead recommend the use of GIC, which does not have any additional parameter.

There are two places where a spatial covariance model is considered. One is for modelling spatial dependence of the data and the other is for constructing a spatial weight matrix in the objective function. Our theory shows that the method is valid for a class of underlying spatial distributions that satisfy condition 1, part (c), and for a class of spatial weight matrices that satisfy condition 1, part (g). However, some spatial covariance models that satisfy condition 1, part (c), may not necessarily satisfy condition 1, part (g). In this regard, our approach is more general as it covers the case that the true spatial covariance matrix is a spatial weight matrix in addition to other spatial covariances as long as both conditions 1, part (c), and 1, part (g), are satisfied.

4. Numerical investigation

4.1. Simulation study

In this section, we present a simulation study to illustrate our theoretical findings. We consider $\mathbf{S}=\{(s_1,s_2), s_i, s_j=1,\ldots,m\}$ with $m=6,12,24$, i.e. sample sizes $n=36,144,576$ respectively. We consider $J=15,25,35$ with four non-zero components ($q=4$) which are $f_1(x)=5x$, $f_2(x)=3(2x-1)^2$, $f_3(x)=4\sin(2\pi x)/\{2-\sin(2\pi x)\}$ and $f_4(x)=0.6\sin(2\pi x)+1.2\cos(2\pi x)+1.8\sin^2(2\pi x)+2.4\cos^3(2\pi x)+3.0\sin^3(2\pi x)$. These non-zero components are taken from Huang *et al.* (2010a) and were originally introduced by Lin and Zhang (2006). The covariates are $X_j=(W_j+tU)/(1+t)$, for $j=1,\ldots,J$, where W_j and U are independently and identically distributed from a uniform[0, 1] distribution. The correlation between X_j and X_k is given as $t^2/(1+t^2)$. This implies that the dependence between covariates increases in t. In the simulation study, we consider $t=1$ and $t=3$.

We assume that the error follows a stationary mean 0 Gaussian process with a spatial covariance function $\delta(\mathbf{h})$. To investigate the selection performance of the method proposed, we consider three different covariance models: exponential, Matérn and Gaussian covariance functions. The exponential and Gaussian covariance functions have two parameters (σ^2 and ρ) and the Matérn covariance function involves one more parameter ν. For simplicity, we set the variance $\sigma^2=1$. For an exponential covariance function, we consider $\rho=0.5$ and $\rho=1$. For a Matérn covariance function, we consider $\nu=\frac{3}{2}$ and $\nu=\frac{5}{2}$ and for each of these cases we set ρ to be 1.5 and 2.5. For a Gaussian covariance function, we consider $\rho=1.5$ and $\rho=2.5$. The parameter ρ contributes to how fast the covariance function decays as the distance $\|\mathbf{h}\|$ increases. We generate 100 data sets for each case.

Three covariance functions are characterized by the mean-square differentiability of the process. A Gaussian process with an exponential covariance function is continuous whereas a Gaussian process with a Gaussian covariance function is infinitely differentiable in a mean-squared sense. As an intermediate, a Gaussian process with a Matérn covariance function is $\lceil \nu-1 \rceil$ times differentiable where $\lceil x \rceil$ is the smallest integer that is larger than or equal to x (Stein, 1999). The mean-square differentiability is related to the smoothness of the processes, i.e. the local behaviour of the processes. Thus, we can also investigate selection performance in view of the local property of the processes by considering different types of covariance function.

For each data set generated, the selection performance of the method proposed is examined under several choices of a spatial weight matrix. In particular, we considered two classes: Gaussian and inverse multiquadratic functions. In addition, we also considered an identity matrix and the true covariance function of the underlying process as a spatial weight matrix for comparison. When a spatial weight matrix is controlled by a spatial weight parameter, we applied the approach that was introduced in Section 3.1 to select both a spatial weight parameter and a penalty parameter. When we consider an identity matrix as the spatial weight matrix, we do not have a spatial weight parameter to control. In this case, we considered a sequence of penalty parameter values around the theoretical lower bound of the penalty parameter and chose the one that minimizes GIC. We also implemented the method of Huang *et al.* (2010a) which was developed for independent data for comparison. We refer to this approach as the 'independence approach'.

Following standard practice, we computed the average and standard deviation of the true positive, TP, and false positive, FP, findings. TP is the number of additive components that are correctly selected and FP is the number of additive components that are falsely selected. In our simulation setting, the desired values of TP and FP are 4 $(=q)$ and 0 respectively. Table 2 shows selected results for the case with $t=1$ when generating X_j, $m=6, 24$ and a selected set of true correlation parameter values. Complete simulation results are given in the on-line supplementary material. To assess our method in the case of increased dependence between covariates, we studied other choices of t, such as $t=3$. Extensive simulation results are available in the supplementary material.

The first row (where the spatial weight is 'none (independence)') in each of the covariance model blocks in Table 2 corresponds to the independence approach. The results clearly indicate overestimation of FP-components. Actually, the independence approach is selecting more components than the truth. The trend continues even for increased sample sizes. The following rows are results from various choices of spatial weight matrices. Our method successfully reduced overestimation. Even when the spatial weight matrix is an identity matrix, our dependent data approach still reduces overestimation compared with the independence approach. The result persists for various sample sizes m, covariance models, choices of spatial weight matrices and the number of covariates, J.

When the true covariance model is exponential or Matérn and we used them (the truth) as a spatial weight matrix, the method proposed did not perform well in terms of TP for small sample sizes. One possible reason is that the exponential and Matérn covariance functions in the spatial weight matrices produced larger maximum eigenvalues of $\mathbf{L}$ for small samples. For example, when $\mathrm{Mat}_{5/2}(2.5)$ is used for a spatial weight, the corresponding maximum eigenvalue of L is 20.32 for $m=6$ whereas the maximum eigenvalue of L is 1.23 for Gauss(1.5) as a spatial weight. A larger maximum eigenvalue of L makes a larger penalty parameter, so fewer components are selected. However, the performance improves as m increases. For small sizes $(m=6)$, inverse multiquadratic spatial weights tend to underestimate so TP is lower compared with other spatial weight matrix choices but improves as m increases. A Gaussian spatial weight matrix maintains a similar level of TP whereas FP is reduced compared with the independence approach. Thus, we recommend the use of a Gaussian spatial weight matrix, in particular for small sample sizes in practice.

Results for increased dependence between covariates (from $t=1$ to $t=3$) show that there is an increase (overestimation) of FP. See the tables in section 6 of the on-line supplementary material. Strong dependence between covariates may hinder the selection power of the variable selection approaches and, in turn, results in selection of more components. However, our approach performs comparatively better than the independence approach.

4.2. Real data example

We consider lung cancer mortality data over the period of 2000–2005, obtained from the 'Surveillance, epidemiology, and end results' (SEER) programme (`www.seer.cancer.gov`) of the

Table 2. Monte Carlo mean (with standard deviations in parentheses) for the selected number of non-zero covariates by using a spatially weighted GL

J	*Covariate model*	*Spatial weight*	*Results for m = 6*		*Results for m = 24*	
			TP	*FP*	*TP*	*FP*
15	Exp(0.5)	None (independence)	3.74 (0.48)	3.63 (1.83)	4 (0)	0.55 (0.77)
		I	3.17 (0.82)	2.02 (1.34)	4 (0)	0 (0)
		Gauss	3.01 (0.85)	1.76 (1.24)	4 (0)	0 (0)
		Inverse multiquadratic	2.94 (0.93)	1.58 (1.44)	4 (0)	0 (0)
		True	1.66 (1.02)	0.29 (0.56)	4 (0)	0 (0)
	$\mathrm{Mat}_{3/2}(2.5)$	None (independence)	3.8 (0.45)	3.63 (1.86)	4 (0)	0.34 (0.57)
		I	3.26 (0.77)	1.87 (1.5)	4 (0)	0.05 (0.22)
		Gauss	2.98 (0.85)	1.64 (1.34)	4 (0)	0.01 (0.1)
		Inverse multiquadratic	2.74 (0.86)	1.5 (1.18)	4 (0)	0 (0)
		True	0.8 (0.68)	0.09 (0.32)	4 (0)	0 (0)
	$\mathrm{Mat}_{5/2}(2.5)$	None (independence)	3.82 (0.41)	3.59 (2.08)	4 (0)	0.58 (0.88)
		I	3.22 (0.76)	2.06 (1.55)	4 (0)	0.04 (0.2)
		Gauss	2.91 (0.89)	1.72 (1.35)	4 (0)	0 (0)
		Inverse multiquadratic	2.97 (0.89)	1.67 (1.33)	4 (0)	0 (0)
		True	0.37 (0.56)	0.04 (0.24)	4 (0)	0 (0)
	Gauss(1.5)	None (independence)	3.76 (0.49)	3.97 (2.06)	4 (0)	0.59 (0.81)
		I	3.23 (0.75)	2.07 (1.24)	4 (0)	0.03 (0.17)
		Gauss	3.02 (0.82)	1.86 (1.25)	4 (0)	0.01 (0.1)
		Inverse multiquadratic	2.76 (0.79)	1.58 (1.17)	4 (0)	0 (0)
		True	3.3 (0.73)	2.32 (1.55)	4 (0)	0.12 (0.33)
35	Exp(0.5)	None (independence)	3.38 (0.74)	6.22 (2.39)	4 (0)	1.49 (1.55)
		I	2.62 (0.94)	3.15 (2.14)	4 (0)	0.04 (0.2)
		Gauss	2.39 (0.98)	2.59 (1.84)	4 (0)	0.02 (0.14)
		Inverse multiquadratic	2.22 (1.04)	2.38 (1.79)	4 (0)	0 (0)
		True	1.21 (0.95)	0.48 (0.78)	4 (0)	0 (0)
	$\mathrm{Mat}_{3/2}(2.5)$	None (independence)	3.48 (0.67)	5.79 (2.32)	4 (0)	1.25 (1.37)
		I	2.71 (0.9)	2.94 (1.75)	4 (0)	0.1 (0.39)
		Gauss	2.57 (0.92)	2.59 (1.6)	4 (0)	0.01 (0.1)
		Inverse multiquadratic	2.25 (0.9)	2.24 (1.68)	4 (0)	0 (0)
		True	0.52 (0.58)	0.12 (0.41)	4 (0)	0 (0)
	$\mathrm{Mat}_{5/2}(2.5)$	None (independence)	3.33 (0.74)	5.59 (2.19)	4 (0)	1.34 (1.51)
		I	2.72 (1)	3.06 (1.97)	4 (0)	0.1 (0.33)
		Gauss	2.53 (1.04)	2.57 (1.81)	4 (0)	0.01 (0.1)
		Inverse multiquadratic	2.25 (0.97)	2.17 (1.68)	4 (0)	0 (0)
		True	0.34 (0.57)	0.05 (0.26)	4 (0)	0 (0)
	Gauss(1.5)	None (independence)	3.15 (0.88)	6.37 (1.95)	4 (0)	1.78 (1.54)
		I	2.44 (1.02)	3.35 (1.98)	4 (0)	0.08 (0.27)
		Gauss	2.24 (0.95)	2.83 (1.61)	4 (0)	0.02 (0.14)
		Inverse multiquadratic	2.18 (0.9)	2.2 (1.41)	4 (0)	0 (0)
		True	2.78 (0.87)	3.66 (1.74)	4 (0)	0.17 (0.45)

National Cancer Institute, USA, as an illustrative example. The SEER data can be accessed by submitting a signed SEER research data agreement (`seer.cancer.gov/data/access.html`). The SEER data include incidence or mortality of cancers and associated variables for US counties. We consider the southern part of Michigan which includes 68 counties. We applied Tukey's transformation (e.g. Cressie and Chan (1989)) to age-adjusted lung cancer mortality rates used as the response variable. We included 20 covariates obtained from the SEER database which were originally from the US Census Bureau. We also added $\mathrm{PM}_{2.5}$ (particulate matter smaller than 2.5 μm in diameter) obtained from the US Environmental Protection Agency's na-

Table 3. Comparing the two methods (independence approach and dependence approach) for the real data example†

Variable	*Independence model*		*Dependence model*	
	GL	*AGL*	*GL*	*AGL*
1	×	×	×	×
2	✓	✓	✓	✓
3	×	×	×	×
4	×	×	×	×
5	×	×	×	×
6	×	×	×	×
7	×	×	×	×
8	✓	✓	×	×
9	×	×	×	×
10	×	×	×	×
11	×	×	×	×
12	×	×	×	×
13	×	×	×	×
14	×	×	×	×
15	×	×	×	×
16	×	×	×	×
17	×	×	×	×
18	×	×	×	×
19	×	×	×	×
20	✓	✓	✓	✓
21	✓	✓	✓	×

†Variable description: 1, population mortality; 2, poverty; 3, PM25; 4, urban; 5, non-white; 6, never married; 7, agriculture; 8, unemployment; 9, white collar; 10, higher high school; 11, age more than 65 years; 12, age less than 18 years; 13, crowding; 14, foreign born; 15, language isolation; 16, median household income; 17, same house no migration; 18, move same county; 19, move same state; 20, move different state; 21, normalized cost of living.

tional emission inventory database (`www.epa.gov/air/data/emisdist.html?st~MI~Michigan`). Since emission data in this Web site are available for the years 2001 and 2002, we considered the average of 2001–2002 emission data. The unit is tons per county.

For our analysis, we scaled each of our predictor variables to [0, 1]. We consider a Gaussian covariance function for the spatial weight matrix and a sequence of ρ-values around the estimated ρ, obtained by fitting an empirical variogram. Then, we applied the selection approach that was introduced in Section 3.1. Selected variables for both GL and AGL algorithms under independent and dependent error are presented in Table 3. Our method of variable selection has a strict sense of selecting variables as it drops more variables. Our approach (the AGL case) dropped two more variables among the variables that were selected by the independence approach. The components selected, ‘poverty’ and ‘move different state’, do not seem to be related to lung cancer mortality rates directly but one may think that poverty can be a proxy for a more relevant covariate for lung cancer mortality rates. For example, a study has shown that smoking is more prevalent in lower socio-economic groups of society (Haustein, 2006). Thus, we can think of poverty as a proxy for tobacco use. Although the variable ‘move different state’ was kept, our approach at least dropped a few more irrelevant variables compared with the independence approach.

To explore further those covariates which were dropped by the approach proposed but selected

Table 4. Summary from linear regression on the selected variables under the independent error assumption

Variable	*Estimate*	*Standard error*	*t-value*	*p-value*
Intercept	0.43958	0.07400	5.940	1.41×10^{-7}
Poverty	0.42467	0.12887	3.295	*0.00163*
Unemployment	0.01086	0.14638	0.074	0.94107
Move same state	−0.01837	0.10624	−0.173	0.86331
Move different state	−0.15474	0.10379	−1.491	*0.14106*
Normalized cost of living	0.02715	0.07346	0.370	0.71298

by the independence approach, we fitted a linear regression model. Although we considered non-linear relationships in spatial additive models in this paper, simple linear regression can provide an initial assessment of the result. We present output from a linear regression model with five covariates in Table 4. This shows that our approach selected two most significant variables on the basis of p-values (italics) whereas the independence approach selected some insignificant variables, such as unemployment and normalized cost of living.

5. Discussion

We established a method of selecting non-zero components in spatial additive models by using a two-step AGL-type approach and a spatial weight in the l_2-error term. The consistency results allow the number of additive components to increase exponentially with the sample size while we assume that the total number of non-zero components is bounded. This setting is common in high dimensional variable selection for additive models (see, for example, Ravikumar *et al.* (2009), Huang *et al.* (2010a) and Lai *et al.* (2012)). Since the lower bound of the penalty parameter depends on a spatial weight matrix, we considered an approach of choosing a spatial weight matrix together with the penalty parameter that works well in practice. Our theoretical result also implies that an identity matrix could be a choice for the spatial weight matrix. This is, however, not an optimal but rather an extreme choice. Nevertheless this performs better than the independent data approach because of a larger theoretical lower bound on the penalty parameter. We provided a guideline on how to select a weight matrix in the real dependent data situation.

One may consider selecting additive components without a spatial weight matrix first and then do non-parametric additive estimation as a second step. Such a two-stage approach could be appropriate for inferential issues instead of selection. In fact, work has begun on this type of two-stage estimation for independent data to reduce the bias (Belloni and Chernozhukov, 2013; Liu and Yu, 2013). Although this is a possible route, the nature of solution is quite different. Unlike just mean estimation, one must control the mean and variance simultaneously.

The condition for the spatial covariance function in assumption 1, part (c), can be extended to $\int_{\mathbf{D}_n}\delta(\mathbf{h})\,\mathrm{d}\mathbf{h}=\mathbf{O}(n^{\alpha})$ for some $\alpha\in[0,1)$. Here $\alpha=0$ corresponds to the current assumption 1, part (c). By allowing $0<\alpha<1$, we can include spatial covariance models of long memory processes, so the theorems under this assumption cover a broader class of spatial covariance functions. The theoretical lower bound of the penalty parameter and the rates of convergence are modified as α is introduced. We provide revised lemmas, theorems and related discussion in the on-line supplementary material for this extended setting.

The theoretical insights that are derived from the upper bounds of the rate of convergence and the lower bound of the penalty parameter may not be a perfect solution. The ideal situation could be finding the rate of convergence as well as the optimal rate for the penalty parameter. Such a task is challenging even in the independent data context and our approach here is in practice optimal in this sense.

Acknowledgements

The authors gratefully acknowledge comments from the reviewers and the Joint Editor that substantially improved the manuscript. The research is partially supported by National Science Foundation grant DMS-1106450.

Appendix A: Proofs of theorems

Before proving the theorems that are stated in Section 3, we introduce some notation and lemmas. Note that, for $f \in \mathcal{F}$, there exists $f_n \in \mathcal{S}_n$ such that $\| f - f_n \|_2 = \mathbf{O}(m_n^{-\tau})$, where $\| \cdot \|_2$ for a function is defined as $\| f \|_2 = \surd \int_a^b f^2(x)\,\mathrm{d}x$ and $\mathcal{S}_n$ is the space that is spanned by B-spline bases (for example see Huang *et al.* (2010a)). Then, define the centred $\mathcal{S}_{nj}^0$ by

$$\mathcal{S}_{nj}^0 = \left\{ f_{nj} : f_{nj} = \sum_{l=1}^{m_n} b_{jl}\, \mathbb{B}_l^{\mathrm{c}}(x), (b_{j1}, \ldots, b_{jm_n}) \in \mathbb{R}^{m_n} \right\}, \qquad 1 \leqslant j \leqslant J. \tag{8}$$

Recall that $\mathbb{B}_l^{\mathrm{c}}(x) = \mathbb{B}_l(x) - (1/n)\Sigma_{\mathbf{s}' \in \mathbf{S}}\, \mathbb{B}_l\{X_j(\mathbf{s}')\}$, which depend on X_j. Also, to emphasize the fact that $\boldsymbol{\pi}$, $\boldsymbol{\epsilon}$ and $\boldsymbol{\theta}$ depend on sample size n we shall use the suffix n for each of these three quantities in our proofs.

Recall that $A_0 = \{ j : f_j(x) \equiv 0, 1 \leqslant j \leqslant J \}$ and $A_* = \{1, 2, \ldots, J\} \setminus A_0$ so that A_0 and A_* are the sets of zero and non-zero components respectively. We introduced $\tilde{A}_0$ as the index set that satisfies $\Sigma_{j \in \tilde{A}_0} ||\boldsymbol{\beta}_j||_2 \leqslant \eta_1$ for some $\eta_1 \geqslant 0$ and let $\tilde{A}_* = \{1, 2, \ldots, J\} \setminus \tilde{A}_0$. Without loss of generality, we can assume that $A_0 \subseteq \tilde{A}_0$ so $\tilde{A}_* \subseteq A_*$ and $q^* := |\tilde{A}_*| \leqslant q$. For any subset $A \subset \{1, 2, \ldots, J\}$, define $\boldsymbol{\beta}_A = (\boldsymbol{\beta}_j, j \in A)'$ and $\boldsymbol{\Omega}_A = \mathbb{D}_A^{\mathrm{c}\prime} \mathbb{D}_A^{\mathrm{c}}/n$, where $\mathbb{D}_A^{\mathrm{c}} = \mathbf{L}' \mathbb{B}_A^{\mathrm{c}}$ and $\mathbb{B}_A^{\mathrm{c}} = (\mathbb{B}_j^{\mathrm{c}}, j \in A)$ is the subdesign matrix that is formed by the columns indexed in the set A. Note that $\boldsymbol{\beta}_{A_0} \equiv \mathbf{0}$. Also, denote the minimum and maximum eigenvalues and the condition number of a matrix $\mathbf{M}$ by $\rho_{\min}(\mathbf{M})$, $\rho_{\max}(\mathbf{M})$ and $\kappa(\mathbf{M})$ respectively.

Lemma 1 (lemma 1 in Huang *et al.* (2010a)). Suppose that $f \in \mathcal{F}$ and $\mathbb{E}\{ f(X_j)\} = 0$. Then, under parts (d) and (e) in assumption 1, there is an $f_n \in \mathcal{S}_{nj}^0$ such that $\| f_n - f \|_2 = \mathbf{O}_p\{ m_n^{-\tau} + \surd(m_n/n)\}$. Particularly, under the choice of $m_n = \mathbf{O}(n^{1/(2\tau+1)})$, we have $\| f_n - f \|_2 = \mathbf{O}_p(m_n^{-\tau}) = \mathbf{O}_p(n^{-\tau/(2\tau+1)})$.

Lemma 2. Suppose that $|A|$ is bounded by a fixed constant independent of n and J. Let $h_n \asymp m_n^{-1}$. Then under parts (d) and (e) in assumption 1, with probability converging to 1,

$$\rho_{\min}(\boldsymbol{\Sigma}_{\mathrm{W}}^{-1}) d_1 h_n \leqslant \rho_{\min}(\boldsymbol{\Omega}_A) \leqslant \rho_{\max}(\boldsymbol{\Omega}_A) \leqslant \rho_{\max}(\boldsymbol{\Sigma}_{\mathrm{W}}^{-1}) d_2 h_n. \tag{9}$$

Additionally, under assumption 1, part (g), result (9) becomes $c_1 h_n \leqslant \rho_{\min}(\boldsymbol{\Omega}_A) \leqslant \rho_{\max}(\boldsymbol{\Omega}_A) \leqslant c_2 h_n$, where d_1, d_2, c_1 and c_2 are some positive constants.

Lemma 3. Define $\mathbf{M}_n$ to be a non-negative-definite matrix of order n, and $T_{jl} = (m_n/n)^{1/2} \mathbf{a}_{jl}' \mathbf{M}_n \boldsymbol{\epsilon}, \forall\, 1 \leqslant j \leqslant J, 1 \leqslant l \leqslant m_n$, where $\mathbf{a}_{jl} = (\mathbb{B}_l^{\mathrm{c}}\{X_j(\mathbf{s})\}, \mathbf{s} \in \mathbf{S})'$ and

$$T_n = \max_{\substack{1 \leqslant j \leqslant J \\ 1 \leqslant l \leqslant m_n}} |T_{jl}|.$$

Then, under parts (d)–(f) in assumption 1, $\mathbb{E}(T_n) \leqslant C_1\, \rho_{\max}(\mathbf{M}_n) \surd\{ m_n \log(J m_n)\}$, for some $C_1 > 0$.

Before we delve into the proof of the theorems, we define and summarize some index sets that we shall be using. Recall that $A_0 = \{ j : f_j \equiv 0, 1 \leqslant j \leqslant J \}$. For an index set $\tilde{A}_1$ that satisfies $\tilde{A}_\beta = \{ j : \| \hat{\boldsymbol{\beta}}_{\mathrm{GL}, j} \|_2 > 0 \} \subseteq \tilde{A}_1 \subseteq \tilde{A}_\beta \cup \tilde{A}_*$, we consider the sets in Table 5.

We can deduce some relationships from Table 5: $\tilde{A}_3 = \tilde{A}_1 \cap \tilde{A}_*$, $\tilde{A}_4 = \tilde{A}_1 \cap \tilde{A}_0$, $\tilde{A}_5 = \tilde{A}_1^{\mathrm{c}} \cap \tilde{A}_*$ and $\tilde{A}_6 = \tilde{A}_2 \cap \tilde{A}_0$, and hence we have $\tilde{A}_3 \cup \tilde{A}_4 = \tilde{A}_1$, $\tilde{A}_5 \cup \tilde{A}_6 = \tilde{A}_2$ and $\tilde{A}_3 \cap \tilde{A}_4 = \tilde{A}_5 \cap \tilde{A}_6 = \phi$. Also, let $|\tilde{A}_1| = q_1$. For an index set $\hat{A}_1$ that satisfies $\hat{A}_f = \{ j : || \hat{f}_{\mathrm{GL}, j} ||_2 > 0 \} \subseteq \hat{A}_1 \subseteq \hat{A}_f \cup A_*$, we have the sets in Table 6.

We have a similar set of relationships from Table 6 which is $\hat{A}_3 = \hat{A}_1 \cap A_*$, $\hat{A}_4 = \hat{A}_1 \cap A_0$, $\hat{A}_5 = \hat{A}_1^{\mathrm{c}} \cap A_*$ and $\hat{A}_6 = \hat{A}_2 \cap A_0$, and hence we have $\hat{A}_3 \cup \hat{A}_4 = \hat{A}_*$, $\hat{A}_5 \cup \hat{A}_6 = \hat{A}_2$ and $\hat{A}_3 \cap \hat{A}_4 = \hat{A}_5 \cap \hat{A}_6 = \phi$.

Table 5

	'Large' $\|\boldsymbol{\beta}_j\|_2$ *(i.e. $\tilde{A}_*$)*	*'Small'* $\|\boldsymbol{\beta}_j\|_2$ *(i.e. $\tilde{A}_0$)*
$\tilde{A}_1$	$\tilde{A}_3$	$\tilde{A}_4$
$\tilde{A}_2=\tilde{A}_1^{\mathrm{c}}$	$\tilde{A}_5$	$\tilde{A}_6$

Table 6

	'Large' f_j *(i.e. A_*)*	*'Small'* f_j *(i.e. A_0)*
$\hat{A}_1$	$\hat{A}_3$	$\hat{A}_4$
$\hat{A}_2=\hat{A}_1^{\mathrm{c}}$	$\hat{A}_5$	$\hat{A}_6$

To prove theorem 1, we need boundedness of $|\tilde{A}_\beta|$, which is given in the following lemma.

Lemma 4. Under assumption 1 with $\lambda_{n1}>C\,\rho_{\max}(\mathbf{L})\surd\{nm_n\log(Jm_n)\}$ for a sufficiently large constant C, we have $|\tilde{A}_\beta|\leqslant M_1|A_*|$ for a finite constant $M_1>1$ with probability converging to 1.

A.1. Proof of theorem 1

A.1.1. Part (a)

Since $\hat{\boldsymbol{\beta}}_{\mathrm{GL}}$ is the GL estimate by minimizing $\mathbf{Q}_{n1}(\boldsymbol{\beta},\lambda_{n1})$, for any $\boldsymbol{\beta}$,

$$\|\mathbf{Z}^{\mathrm{c}}-\mathbb{D}^{\mathrm{c}}\hat{\boldsymbol{\beta}}_{\mathrm{GL}}\|_2^2+\lambda_{n1}\|\hat{\boldsymbol{\beta}}_{\mathrm{GL}}\|_{2,1,\mathbf{1}}\leqslant\|\mathbf{Z}^{\mathrm{c}}-\mathbb{D}^{\mathrm{c}}\boldsymbol{\beta}\|_2^2+\lambda_{n1}\|\boldsymbol{\beta}|_{2,1,\mathbf{1}}. \tag{10}$$

Let $A_2=\{j:\|\boldsymbol{\beta}_j\|_2>0 \text{ or } \|\hat{\boldsymbol{\beta}}_{\mathrm{GL},j}\|_2>0\}$, where $\boldsymbol{\beta}_{A_2}=(\boldsymbol{\beta}_j,\,j\in A_2)$ and $\hat{\boldsymbol{\beta}}_{\mathrm{GL},A_2}=(\hat{\boldsymbol{\beta}}_{\mathrm{GL},j},\,j\in A_2)$. Recall that $\|\boldsymbol{\beta}_{A_2}\|_{2,1,\mathbf{1}}=\Sigma_{j\in A_2}\|\boldsymbol{\beta}_j\|_2$ and $\|\hat{\boldsymbol{\beta}}_{\mathrm{GL},A_2}\|_{2,1,\mathbf{1}}=\Sigma_{j\in A_2}\|\hat{\boldsymbol{\beta}}_{\mathrm{GL},j}\|_2$. By result (10), we have

$$\|\mathbf{Z}^{\mathrm{c}}-\mathbb{D}^{\mathrm{c}}_{A_2}\hat{\boldsymbol{\beta}}_{\mathrm{GL},A_2}\|_2^2+\lambda_{n1}\|\hat{\boldsymbol{\beta}}_{\mathrm{GL},A_2}\|_{2,1,\mathbf{1}}\leqslant\|\mathbf{Z}^{\mathrm{c}}-\mathbb{D}^{\mathrm{c}}_{A_2}\boldsymbol{\beta}_{A_2}\|_2^2+\lambda_{n1}\|\boldsymbol{\beta}_{A_2}\|_{2,1,\mathbf{1}}. \tag{11}$$

Let $\boldsymbol{\vartheta}_{A_2}=\mathbf{Z}^{\mathrm{c}}-\mathbb{D}^{\mathrm{c}}_{A_2}\boldsymbol{\beta}_{A_2}$ and $\boldsymbol{\zeta}_{A_2}=\mathbb{D}^{\mathrm{c}}_{A_2}(\hat{\boldsymbol{\beta}}_{\mathrm{GL},A_2}-\boldsymbol{\beta}_{A_2})$. Using the fact that $\|\mathbf{a}-\mathbf{b}\|_2^2=\|\mathbf{a}\|_2^2-2\mathbf{a}'\mathbf{b}+\|\mathbf{b}\|_2^2$ and $\mathbf{Z}^{\mathrm{c}}-\mathbb{D}^{\mathrm{c}}_{A_2}\hat{\boldsymbol{\beta}}_{\mathrm{GL},A_2}=\mathbf{Z}^{\mathrm{c}}-\mathbb{D}^{\mathrm{c}}_{A_2}\boldsymbol{\beta}_{A_2}-\mathbb{D}^{\mathrm{c}}_{A_2}(\hat{\boldsymbol{\beta}}_{\mathrm{GL},A_2}-\boldsymbol{\beta}_{A_2})$, we can rewrite inequality (11) such that

$$\|\boldsymbol{\zeta}_{A_2}\|_2^2-2\boldsymbol{\vartheta}'_{A_2}\boldsymbol{\zeta}_{A_2}\leqslant\lambda_{n1}\|\boldsymbol{\beta}_{A_2}\|_{2,1,\mathbf{1}}-\lambda_{n1}\|\hat{\boldsymbol{\beta}}_{\mathrm{GL},A_2}\|_{2,1,\mathbf{1}}\leqslant\lambda_{n1}\surd|A_2|\,\|\hat{\boldsymbol{\beta}}_{\mathrm{GL},A_2}-\boldsymbol{\beta}_{A_2}\|_2. \tag{12}$$

Subsequent steps will be to bound $\|\hat{\boldsymbol{\beta}}_{\mathrm{GL},A_2}-\boldsymbol{\beta}_{A_2}\|_2^2$. First, we have

$$2|\boldsymbol{\vartheta}'_{A_2}\boldsymbol{\zeta}_{A_2}|\leqslant2\|\boldsymbol{\vartheta}^*_{A_2}\|_2\|\boldsymbol{\zeta}_{A_2}\|_2=2(\surd2\|\boldsymbol{\vartheta}^*_{A_2}\|_2)\frac{\|\boldsymbol{\zeta}_{A_2}\|_2}{\surd2}\leqslant2\|\boldsymbol{\vartheta}^*_{A_2}\|_2^2+\frac{\|\boldsymbol{\zeta}_{A_2}\|_2^2}{2}, \tag{13}$$

where the last inequality is based on the fact that $2ab\leqslant a^2+b^2$ and $\boldsymbol{\vartheta}^*_{A_2}$ is the projection of $\boldsymbol{\vartheta}_{A_2}$ onto the span of $\mathbb{D}^{\mathrm{c}}_{A_2}$. Now, by combining expressions (12) and (13), we have

$$\|\boldsymbol{\zeta}_{A_2}\|_2^2\leqslant4\|\boldsymbol{\vartheta}^*_{A_2}\|_2^2+2\lambda_{n1}\surd|A_2|\,\|\hat{\boldsymbol{\beta}}_{\mathrm{GL},A_2}-\boldsymbol{\beta}_{A_2}\|_2. \tag{14}$$

However, we have

$$\begin{aligned}\|\boldsymbol{\zeta}_{A_2}\|_2^2&=(\hat{\boldsymbol{\beta}}_{\mathrm{GL},A_2}-\boldsymbol{\beta}_{A_2})'\mathbb{D}^{\mathrm{c}\prime}_{A_2}\mathbb{D}^{\mathrm{c}}_{A_2}(\hat{\boldsymbol{\beta}}_{\mathrm{GL},A_2}-\boldsymbol{\beta}_{A_2})\geqslant n\,\rho_{\min}\left(\frac{\mathbb{D}^{\mathrm{c}\prime}_{A_2}\mathbb{D}^{\mathrm{c}}_{A_2}}{n}\right)\|\hat{\boldsymbol{\beta}}_{\mathrm{GL},A_2}-\boldsymbol{\beta}_{A_2}\|_2^2\\&=n\,\rho_{\min}(\boldsymbol{\Omega}_{A_2})\|\hat{\boldsymbol{\beta}}_{\mathrm{GL},A_2}-\boldsymbol{\beta}_{A_2}\|_2^2\geqslant nc_1h_n\|\hat{\boldsymbol{\beta}}_{\mathrm{GL},A_2}-\boldsymbol{\beta}_{A_2}\|_2^2,\end{aligned} \tag{15}$$

where the last inequality is from lemma 2. Combining expressions (14) and (15), we have

$$\begin{aligned} nc_1h_n\|\hat{\boldsymbol{\beta}}_{\mathrm{GL},A_2}-\boldsymbol{\beta}_{A_2}\|_2^2 &\leqslant 4\|\boldsymbol{\vartheta}^*_{A_2}\|_2^2+2\lambda_{n1}\surd|A_2|\|\boldsymbol{\beta}_{A_2}-\hat{\boldsymbol{\beta}}_{\mathrm{GL},A_2}\|_2 \\ &=4\|\boldsymbol{\vartheta}^*_{A_2}\|_2^2+2\frac{\lambda_{n1}\surd(2|A_2|)}{\surd(nc_1h_n)}\left\{\|\boldsymbol{\beta}_{A_2}-\hat{\boldsymbol{\beta}}_{\mathrm{GL},A_2}\|_2\sqrt{\left(\frac{nc_1h_n}{2}\right)}\right\} \\ &\leqslant 4\|\boldsymbol{\vartheta}^*_{A_2}\|_2^2+\frac{2\lambda_{n1}^2|A_2|}{nc_1h_n}+\|\boldsymbol{\beta}_{A_2}-\hat{\boldsymbol{\beta}}_{\mathrm{GL},A_2}\|_2^2\frac{nc_1h_n}{2} \end{aligned}$$

so that

$$\|\hat{\boldsymbol{\beta}}_{\mathrm{GL},A_2}-\boldsymbol{\beta}_{A_2}\|_2^2\leqslant\frac{8\|\boldsymbol{\vartheta}^*_{A_2}\|_2^2}{nc_1h_n}+\frac{4\lambda_{n1}^2|A_2|}{n^2c_1^2h_n^2}. \tag{16}$$

Let $\vartheta(\mathbf{s})$ be the entry of $\boldsymbol{\vartheta}_{A_2}$. Then, we can rewrite $\vartheta(\mathbf{s})=\epsilon(\mathbf{s})+\mu-\bar{Y}+\mathbf{f}_0\{X(\mathbf{s})\}-\mathbf{f}_{nA_2}\{X(\mathbf{s})\}$, where $\mathbf{f}_0\{X(\mathbf{s})\}=\Sigma_{j=1}^J\,\mathrm{f}_j\{X_j(\mathbf{s})\}$ and $\mathbf{f}_{nA_2}\{X(\mathbf{s})\}=\Sigma_{j\in A_2}\,\mathrm{f}_{nj}\{X_j(\mathbf{s})\}$. Note that $|\mu-\bar{Y}|^2=\mathbf{O}_p(n^{-1})$. Let $\boldsymbol{\epsilon}^*_{A_2}$ be the projection of $\boldsymbol{\epsilon}_n$ onto the span of $\mathbb{D}^{\mathrm{c}}_{A_2}$, i.e. $\boldsymbol{\epsilon}^*_{A_2}=(\mathbb{D}^{\mathrm{c}\prime}_{A_2}\mathbb{D}^{\mathrm{c}}_{A_2})^{-1/2}\mathbb{D}^{\mathrm{c}\prime}_{A_2}\boldsymbol{\epsilon}_n$. Then, we have

$$\begin{aligned} \|\boldsymbol{\vartheta}^*_{A_2}\|_2^2 &\leqslant \|\boldsymbol{\epsilon}^*_{A_2}\|_2^2+\mathbf{O}_p(1+|A_2|nm_n^{-2\tau}) &&(17)\\ &=\|(\mathbb{D}^{\mathrm{c}\prime}_{A_2}\mathbb{D}^{\mathrm{c}}_{A_2})^{-1/2}\mathbb{D}^{\mathrm{c}\prime}_{A_2}\boldsymbol{\epsilon}_n\|_2^2+\mathbf{O}_p(1+|A_2|nm_n^{-2\tau}) \\ &\leqslant\frac{\|\mathbb{D}^{\mathrm{c}\prime}_{A_2}\boldsymbol{\epsilon}_n\|_2^2}{nc_1h_n}+\mathbf{O}_p(1+|A_2|nm_n^{-2\tau}) \qquad \text{(by lemma 2)} \\ &\leqslant\frac{1}{nc_1h_n}\max_{A:|A|\leqslant|A_2|}\|\mathbb{D}^{\mathrm{c}\prime}_{A}\boldsymbol{\epsilon}_n\|_2^2+\mathbf{O}_p(1+|A_2|nm_n^{-2\tau}) \\ &\leqslant\frac{1}{nc_1h_n}|A_2|m_n\max_{1\leqslant j\leqslant J,1\leqslant l\leqslant m_n}\left|\mathbb{D}^{\mathrm{c}\prime}_{jl}\boldsymbol{\epsilon}_n\right|^2+\mathbf{O}_p(1+|A_2|nm_n^{-2\tau}) \\ &=\frac{1}{c_1h_n}|A_2|\max_{1\leqslant j\leqslant J,1\leqslant l\leqslant m_n}\left|\left(\frac{m_n}{n}\right)^{1/2}\mathbf{a}'_{jl}\mathbf{L}\boldsymbol{\epsilon}_n\right|^2+\mathbf{O}_p(1+|A_2|nm_n^{-2\tau}) \\ &=\mathbf{O}_p\left\{\frac{|A_2|\rho^2_{\max}(\mathbf{L})m_n\log(Jm_n)}{c_1h_n}\right\}+\mathbf{O}_p(1+|A_2|nm_n^{-2\tau}), &&(18) \end{aligned}$$

where the last equality is by lemma 3 using $\mathbf{M}_n=\mathbf{L}$. Part (a) of theorem 1 follows by combining expressions (16) and (18) since $|A_2|$ is bounded by lemma 4.

A.1.2. Part (b)

If $m_n^2\lambda_{n1}^2/n^2\rightarrow 0$ then, by the condition $\lambda_{n1}>C\,\rho_{\max}(\mathbf{L})\surd\{nm_n\log(Jm_n)\}$, we have $\rho^2_{\max}(\mathbf{L})m_n^3\log(Jm_n)/n\rightarrow 0$, also note that $1=\rho_{\max}(\mathbb{I})=\rho_{\max}(\boldsymbol{\Sigma}_{\mathrm{W}}^{-1}\boldsymbol{\Sigma}_{\mathrm{W}})\leqslant\rho_{\max}(\boldsymbol{\Sigma}_{\mathrm{W}}^{-1})\rho_{\max}(\boldsymbol{\Sigma}_{\mathrm{W}})\leqslant\rho^2_{\max}(\mathbf{L})\,\rho_{\max}(\boldsymbol{\Sigma}_{\mathrm{W}})$ therefore $\rho^2_{\max}(\mathbf{L})\geqslant\rho_{\min}(\boldsymbol{\Sigma}_{\mathrm{W}}^{-1})$ and hence

$$\begin{aligned} \frac{\rho^2_{\max}(\mathbf{L})m_n^3\log(Jm_n)}{n}&=\rho^2_{\max}(\mathbf{L})m_n^2\log(Jm_n)\frac{m_n}{n}\geqslant\rho_{\min}(\boldsymbol{\Sigma}_{\mathrm{W}}^{-1})m_n^2\log(Jm_n)\frac{m_n}{n} \\ &\geqslant Cm_n^2\log(Jm_n)\frac{m_n}{n}, \end{aligned} \tag{19}$$

where C is a generic constant. Since the left-hand side of inequality (19) goes to 0 and $m_n^2\log(Jm_n)\rightarrow\infty$, $m_n/n\rightarrow 0$. Similarly,

$$\begin{aligned} \frac{\rho^2_{\max}(\mathbf{L})m_n^3\log(Jm_n)}{n}&=\frac{\rho^2_{\max}(\mathbf{L})m_n^{2\tau+2}\log(Jm_n)}{nm_n^{2\tau-1}}=\frac{\rho_{\min}(\boldsymbol{\Sigma}_{\mathrm{W}}^{-1})n^{(2\tau+2)/(2\tau-1)}\log(Jm_n)}{nm_n^{2\tau-1}} \\ &\geqslant\rho_{\min}(\boldsymbol{\Sigma}_{\mathrm{W}}^{-1})\log(Jm_n)\frac{1}{m_n^{2\tau-1}}\geqslant C\log(Jm_n)\frac{1}{m_n^{2\tau-1}}. \end{aligned} \tag{20}$$

Since the left-hand side of inequality (20) goes to 0 and $\log(Jm_n)\rightarrow\infty$, $1/m_n^{2\tau-1}\rightarrow 0$. Thus, we have part (b) of theorem 1.

A.2. Proof of theorem 2

Part (a) of theorem 2 is from the fact that $c_* m_n^{-1}\|\hat{\boldsymbol{\beta}}_{\mathrm{GL},j}-\boldsymbol{\beta}_j\|_2 \leqslant \|\hat{f}_{\mathrm{GL},j}-f_j\|_2 \leqslant c^* m_n^{-1}\|\hat{\boldsymbol{\beta}}_{\mathrm{GL},j}-\boldsymbol{\beta}_j\|_2$ for some $c_*, c^* > 0$. Part (b) follows from part (a).

A.3. Proof of theorem 3

A.3.1. Part (a)

Recall that, by the Karush–Kuhn–Tucker conditions, a necessary and sufficient condition for $\hat{\boldsymbol{\beta}}_{\mathrm{AGL}}$ is

$$\begin{aligned} \mathbb{D}_j^{\mathrm{c}\prime}(\mathbf{Z}^{\mathrm{c}}-\mathbb{D}^{\mathrm{c}}\hat{\boldsymbol{\beta}}_{\mathrm{AGL}}) &= \lambda_{n2}\eta_{nj}\frac{\hat{\boldsymbol{\beta}}_{\mathrm{AGL},j}}{2\|\hat{\boldsymbol{\beta}}_{\mathrm{AGL},j}\|_2}, \qquad \text{when } \|\hat{\boldsymbol{\beta}}_{\mathrm{AGL},j}\|_2>0, \\ \|\mathbb{D}_j^{\mathrm{c}\prime}(\mathbf{Z}^{\mathrm{c}}-\mathbb{D}^{\mathrm{c}}\hat{\boldsymbol{\beta}}_{\mathrm{AGL}})\|_2 &\leqslant \frac{\lambda_{n2}\eta_{nj}}{2}, \qquad \text{when } \|\hat{\boldsymbol{\beta}}_{\mathrm{AGL},j}\|_2=0. \end{aligned} \tag{21}$$

Let $A_{**}=A_*\cap\{j:\|\hat{\boldsymbol{\beta}}_{\mathrm{AGL},j}\|_2>0\}$. Define

$$\hat{\boldsymbol{\beta}}_{A_{**}}=(\mathbb{D}_{A_{**}}^{\mathrm{c}\prime}\mathbb{D}_{A_{**}}^{\mathrm{c}})^{-1}(\mathbb{D}_{A_{**}}^{\mathrm{c}\prime}\mathbf{Z}^{\mathrm{c}}-\lambda_{n2}\mathbf{v}_n), \tag{22}$$

where $\mathbf{v}_n=(v_{nj}, j\in A_{**})$ with $v_{nj}=\eta_{nj}\hat{\boldsymbol{\beta}}_j/(2\|\hat{\boldsymbol{\beta}}_j\|_2)$. Then, we have

$$\mathbb{D}_j^{\mathrm{c}\prime}(\mathbf{Z}^{\mathrm{c}}-\mathbb{D}_{A_{**}}^{\mathrm{c}}\hat{\boldsymbol{\beta}}_{A_{**}})=\lambda_{n2}\eta_{nj}\hat{\boldsymbol{\beta}}_j/(2\|\hat{\boldsymbol{\beta}}_j\|_2),$$

for $j\in A_{**}$. If we assume that $\|\mathbb{D}_j^{\mathrm{c}\prime}(\mathbf{Z}^{\mathrm{c}}-\mathbb{D}_{A_{**}}^{\mathrm{c}}\hat{\boldsymbol{\beta}}_{A_{**}})\|_2\leqslant\lambda_{n2}\eta_{nj}/2$ for all $j\notin A_{**}$, then condition (21) holds for $(\hat{\boldsymbol{\beta}}_{A_{**}}', \mathbf{0}')$, so $\hat{\boldsymbol{\beta}}_{\mathrm{AGL}}=(\hat{\boldsymbol{\beta}}_{A_{**}}', \mathbf{0}')$ since $\mathbb{D}^{\mathrm{c}}\hat{\boldsymbol{\beta}}=\mathbb{D}_{A_{**}}^{\mathrm{c}}\hat{\boldsymbol{\beta}}_{A_{**}}$. If $\|\boldsymbol{\beta}_j\|_2-\|\hat{\boldsymbol{\beta}}_j\|_2<\|\boldsymbol{\beta}_j\|_2$ for all $j\in A_{**}$, then $\hat{\boldsymbol{\beta}}_{A_{**}}=_0\boldsymbol{\beta}_{A_{**}}$ so we have $\hat{\boldsymbol{\beta}}_{\mathrm{AGL}}=_0\boldsymbol{\beta}$.

Therefore, we can have the following inequalities:

$$\begin{aligned} \mathbb{P}(\hat{\boldsymbol{\beta}}_{\mathrm{AGL}}\neq_0\boldsymbol{\beta}) &\leqslant \mathbb{P}(\|\boldsymbol{\beta}_j\|_2-\|\hat{\boldsymbol{\beta}}_j\|_2\geqslant\|\boldsymbol{\beta}_j\|_2, \exists\, j\in A_{**}) \\ &\quad +\mathbb{P}\{\|\mathbb{D}_j^{\mathrm{c}\prime}(\mathbf{Z}-\mathbb{D}_{A_{**}}^{\mathrm{c}}\hat{\boldsymbol{\beta}}_{A_{**}})\|_2>\lambda_{n2}\eta_{nj}/2, \exists\, j\notin A_{**}\} \\ &\leqslant \mathbb{P}(\|\hat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j\|_2\geqslant\|\boldsymbol{\beta}_j\|_2, \exists\, j\in A_{**}) \\ &\quad +\mathbb{P}\{\|\mathbb{D}_j^{\mathrm{c}\prime}(\mathbf{Z}-\mathbb{D}_{A_*}^{\mathrm{c}}\hat{\boldsymbol{\beta}}_{A_{**}})\|_2>\lambda_{n2}\eta_{nj}/2, \exists\, j\notin A_{**}\}, \end{aligned} \tag{23}$$

where the last inequality is from $\|\boldsymbol{\beta}_j\|_2>0$ for $j\in A_{**}$. First, we show that

$$\mathbb{P}(\|\hat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j\|_2\geqslant\|\boldsymbol{\beta}_j\|_2, \exists\, j\in A_{**})\rightarrow 0. \tag{24}$$

To show result (24), it is sufficient to show that $\max_{j\in A_{**}}\|\hat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j\|_2\rightarrow 0$ in probability since $\|\boldsymbol{\beta}_j\|_2>0$ for $j\in A_{**}$. Define $\mathbf{T}_{nj}=(\mathbb{O}_{m_n},\ldots,\mathbb{O}_{m_n},\mathbb{I}_{m_n},\mathbb{O}_{m_n},\ldots,\mathbb{O}_{m_n})$ to be an $m_n\times qm_n$ matrix with $\mathbb{I}_{m_n}$ in the jth block, $\mathbb{O}_{m_n}$ is an $m_n\times m_n$ matrix of 0s and $\mathbb{I}_{m_n}$ is an $m_n\times m_n$ identity matrix. From equation (22), $\hat{\boldsymbol{\beta}}_{A_{**}}-\boldsymbol{\beta}_{A_{**}}=n^{-1}\boldsymbol{\Omega}_{A_{**}}^{-1}(\mathbb{D}_{A_{**}}^{c\prime}\boldsymbol{\epsilon}_n+\mathbb{D}_{A_{**}}^{c\prime}\boldsymbol{\theta}_n-\lambda_{n2}\mathbf{v}_n)$. Thus, if $j\in A_{**}$, we have $\hat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j=n^{-1}\mathbf{T}_{nj}\boldsymbol{\Omega}_{A_{**}}^{-1}(\mathbb{D}_{A_{**}}^{c\prime}\boldsymbol{\epsilon}_n+\mathbb{D}_{A_{**}}^{c\prime}\theta_n-\lambda_{n2}\mathbf{v}_n)$. By triangle inequality,

$$\|\hat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j\|_2\leqslant\frac{\|\mathbf{T}_{nj}\boldsymbol{\Omega}_{A_{**}}^{-1}\mathbb{D}_{A_{**}}^{\mathrm{c}\prime}\boldsymbol{\epsilon}_n\|_2}{n}+\frac{\|\mathbf{T}_{nj}\boldsymbol{\Omega}_{A_{**}}^{-1}\mathbb{D}_{A_{**}}^{\mathrm{c}\prime}\boldsymbol{\theta}_n\|_2}{n}+\frac{\lambda_{n2}\|\mathbf{T}_{nj}\boldsymbol{\Omega}_{A_{**}}^{-1}\mathbf{v}_n\|_2}{n}. \tag{25}$$

We show that each term on the right-hand side in inequality (25) goes to 0 in probability. For the first term,

$$\begin{aligned} \max_{j\in A_{**}}\frac{\|\mathbf{T}_{nj}\boldsymbol{\Omega}_{A_{**}}^{-1}\mathbb{D}_{A_{**}}^{\mathrm{c}\prime}\boldsymbol{\epsilon}_n\|_2}{n} &\leqslant\frac{\|\mathbb{D}_{A_{**}}^{\mathrm{c}\prime}\boldsymbol{\epsilon}_n\|_2}{n\rho_{\max}(\boldsymbol{\Omega}_{A_{**}})}\leqslant\frac{\sqrt{|A_{**}|}}{n^{1/2}\rho_{\max}(\boldsymbol{\Omega}_{A_{**}})}\sqrt{\left(\max_{\substack{j\in A_{**}\\ 1\leqslant l\leqslant m_n}}\frac{m_n}{n}|\mathbb{D}_{jl}^{\mathrm{c}\,\prime}\boldsymbol{\epsilon}_n|^2\right)} \\ &=\mathbf{O}_p\left[\sqrt{\left\{\frac{\rho_{\max}^2(\mathbf{L})m_n^3\log(|A_{**}|m_n)}{n}\right\}}\right]\rightarrow 0 \end{aligned} \tag{26}$$

where the last equality holds by lemma 3 and results (26) holds by assumptions 1, parts (f) and (g). For the second term,

$$\begin{aligned} \max_{j\in A_{**}}n^{-1}\|\mathbf{T}_{nj}\boldsymbol{\Omega}_{A_{**}}^{-1}\mathbb{D}_{A_{**}}^{\mathrm{c}\prime}\boldsymbol{\theta}_n\|_2 &\leqslant n^{-1/2}\|\boldsymbol{\Omega}_{A_{**}}^{-1}\|_2\|n^{-1}\mathbb{D}_{A_{**}}^{\mathrm{c}\prime}\mathbb{D}_{A_{**}}^{\mathrm{c}}\|_2^{1/2}\|\boldsymbol{\theta}_n\|_2 \\ &\leqslant n^{-1/2}\rho_{\min}^{-1}(\boldsymbol{\Omega}_{A_{**}})\rho_{\max}^{1/2}(\boldsymbol{\Omega}_{A_{**}})\mathbf{O}_p(n^{1/(4\tau+2)})=\mathbf{O}_p(n^{1/(2\tau+1)-1/2})\rightarrow 0, \end{aligned} \tag{27}$$

where result (27) holds by assumption 1, part (f). For the third term, we first find an upper bound for $\|\mathbf{v}_n\|_2^2$:

$$\begin{aligned}\|\mathbf{v}_n\|_2^2 &= \frac{1}{2}\sum_{j\in A_{**}} \eta_{nj}^2 = \frac{1}{2}\sum_{j\in A_{**}} \|\hat{\boldsymbol{\beta}}_{\mathrm{GL},j}\|_2^{-2} = \frac{1}{2}\sum_{j\in A_{**}} \frac{\|\boldsymbol{\beta}_j\|_2^2 - \|\hat{\boldsymbol{\beta}}_{\mathrm{GL},j}\|_2^2 + \|\hat{\boldsymbol{\beta}}_{\mathrm{GL},j}\|_2^2}{\|\hat{\boldsymbol{\beta}}_{\mathrm{GL},j}\|_2^2 \|\boldsymbol{\beta}_j\|_2^2} \\ &= \frac{1}{2}\sum_{j\in A_{**}} \frac{\|\boldsymbol{\beta}_j\|_2^2 - \|\hat{\boldsymbol{\beta}}_{\mathrm{GL},j}\|_2^2}{\|\hat{\boldsymbol{\beta}}_{\mathrm{GL},j}\|_2^2 \|\boldsymbol{\beta}_j\|_2^2} + \frac{1}{2}\sum_{j\in A_{**}} \|\boldsymbol{\beta}_j\|_2^{-2} \leqslant Ck_b^{-2} b_{n1}^{-4} \|\hat{\boldsymbol{\beta}}_{\mathrm{GL},A_{**}} - \boldsymbol{\beta}_{A_{**}}\|_2^2 + qb_{n1}^{-2} \\ &= \mathbf{O}_p(k_b^{-2} b_{n1}^{-4} r_n^{-1} + qb_{n1}^{-2}) = \mathbf{O}_p(k_n^2), \end{aligned} \tag{28}$$

where C is a generic constant. Then, we have

$$\begin{aligned}\max_{j\in A_{**}} n^{-1}\lambda_{n2} \|\mathbf{T}_{nj} \boldsymbol{\Omega}_{A_{**}}^{-1} \mathbf{v}_n\|_2 &\leqslant n^{-1}\lambda_{n2}\, \rho_{\min}^{-1}(\boldsymbol{\Omega}_{A_{**}}) \|\mathbf{v}_n\|_2 = \mathbf{O}_p\{n^{-1}\lambda_{n2}\, \rho_{\min}^{-1}(\boldsymbol{\Omega}_{A_{**}}) k_n\} \\ &= \mathbf{O}_p\{n^{-1}\lambda_{n2}(r_n^{-1/2} + m_n^{1/2})\} = \mathbf{O}_p\left(\frac{\lambda_{n2} m_n^{1/2}}{n}\right) \to 0, \end{aligned} \tag{29}$$

where result (29) is implied by assumption 2, part (b). Therefore by combining result (26), (27) and (29), we have result (24). Now, we show that

$$\mathbb{P}\{\|\mathbb{D}_j^{\mathrm{c}\prime}(\mathbf{Z}^{\mathrm{c}} - \mathbb{D}_{A_*}^{\mathrm{c}} \hat{\boldsymbol{\beta}}_{A_{**}})\|_2 > \lambda_{n2}\eta_{nj}/2, \exists\, j \notin A_{**}\} \to 0. \tag{30}$$

As $\boldsymbol{\eta}_{nj} = \|\hat{\boldsymbol{\beta}}_{\mathrm{GL},j}\|_2^{-1} = \mathbf{O}_p(r_n)$ for $j \notin A_{**}$, instead of result (30) it is sufficient to show that

$$\mathbb{P}\{\|\mathbb{D}_j^{\mathrm{c}\prime}(\mathbf{Z}^{\mathrm{c}} - \mathbb{D}_{A_{**}}^{\mathrm{c}} \hat{\boldsymbol{\beta}}_{A_{**}})\|_2 > \lambda_{n2} r_n/2, \exists\, j \notin A_{**}\} \to 0. \tag{31}$$

For $j \notin A_{**}$,

$$\begin{aligned}\mathbb{D}_j^{\mathrm{c}\prime}(\mathbf{Z}^{\mathrm{c}} - \mathbb{D}_{A_{**}}^{\mathrm{c}} \hat{\boldsymbol{\beta}}_{A_{**}}) &= \mathbb{D}_j^{\mathrm{c}\prime}\{\mathbf{Z}^{\mathrm{c}} - \mathbb{D}_{A_{**}}^{\mathrm{c}}(\mathbb{D}_{A_{**}}^{\mathrm{c}\prime} \mathbb{D}_{A_{**}}^{\mathrm{c}})^{-1} \mathbb{D}_{A_{**}}^{\mathrm{c}\prime} \mathbf{Z}^{\mathrm{c}} + \lambda_{n2} n^{-1} \mathbb{D}_{A_{**}}^{\mathrm{c}} \boldsymbol{\Omega}_{A_{**}}^{-1} \mathbf{v}_n\} \\ &= \mathbb{D}_j^{\mathrm{c}\prime} \mathbf{H}_n \mathbf{Z}^{\mathrm{c}} + \lambda_{n2} n^{-1} \mathbb{D}_j^{\mathrm{c}\prime} \mathbb{D}_{A_{**}}^{\mathrm{c}} \boldsymbol{\Omega}_{A_{**}}^{-1} v_n \\ &= \mathbb{D}_j^{\mathrm{c}\prime} \mathbf{H}_n \mathbb{D}^{\mathrm{c}} \beta + \mathbb{D}_j^{\mathrm{c}\prime} \mathbf{H}_n \boldsymbol{\epsilon}_n + \mathbb{D}_j^{\mathrm{c}\prime} \mathbf{H}_n \boldsymbol{\theta}_n + \lambda_{n2} n^{-1} \mathbb{D}_j^{\mathrm{c}\prime} \mathbb{D}_{A_{**}}^{\mathrm{c}} \boldsymbol{\Omega}_{A_{**}}^{-1} \mathbf{v}_n \\ &= \mathbb{D}_j^{\mathrm{c}\prime} \mathbb{D}_{A_{**}^{\mathrm{c}}}^{\mathrm{c}} \boldsymbol{\beta}_{A_{**}^{\mathrm{c}}} + \mathbb{D}_j^{\mathrm{c}\prime} \mathbf{H}_n \boldsymbol{\epsilon}_n + \mathbb{D}_j^{\mathrm{c}\prime} \mathbf{H}_n \boldsymbol{\theta}_n + \lambda_{n2} n^{-1} \mathbb{D}_j^{\mathrm{c}\prime} \mathbb{D}_{A_{**}}^{\mathrm{c}} \boldsymbol{\Omega}_{A_{**}}^{-1} \mathbf{v}_n \\ &= \mathbb{D}_j^{\mathrm{c}\prime} \mathbb{D}_{A_* \cap A_{**}^{\mathrm{c}}}^{\mathrm{c}} \boldsymbol{\beta}_{A_* \cap A_{**}^{\mathrm{c}}} + \mathbb{D}_j^{\mathrm{c}\prime} \mathbf{H}_n \boldsymbol{\epsilon}_n + \mathbb{D}_j^{\mathrm{c}\prime} \mathbf{H}_n \boldsymbol{\theta}_n + \lambda_{n2} n^{-1} \mathbb{D}_j^{\mathrm{c}\prime} \mathbb{D}_{A_{**}}^{\mathrm{c}} \boldsymbol{\Omega}_{A_*}^{-1} \mathbf{v}_n, \end{aligned} \tag{32}$$

where $\mathbf{H}_n = \mathbb{I} - P_{A_{**}}$, the first equality is obtained by replacing $\hat{\boldsymbol{\beta}}_{A_{**}}$ with expression (22) and the fourth equality is obtained because $\mathbf{H}_n$ is the projection matrix onto A_{**}^{c}.

By equation (32), the left-hand side of expression (31) can be bounded above by

$$\begin{aligned}\mathbb{P}\{\|\mathbb{D}_j^{\mathrm{c}\prime}(\mathbf{Z}^{\mathrm{c}} - \mathbb{D}_{A_{**}}^{\mathrm{c}} \hat{\boldsymbol{\beta}}_{\mathrm{AGL},A_{**}})\|_2 > \lambda_{n2} r_n/2, \exists\, j \notin A_{**}\} &\leqslant \mathbb{P}(\|\mathbb{D}_j^{\mathrm{c}\prime} \mathbb{D}_{A_* \cap A_{**}^{\mathrm{c}}}^{\mathrm{c}} \boldsymbol{\beta}_{A_* \cap A_{**}^{\mathrm{c}}}\|_2 > \lambda_{n2} r_n/8, \exists\, j \notin A_{**}) \\ &\quad + \mathbb{P}(\|\mathbb{D}_j^{\mathrm{c}\prime} \mathbf{H}_n \boldsymbol{\epsilon}_n\|_2 > \lambda_{n2} r_n/8, \exists\, j \notin A_{**}) \\ &\quad + \mathbb{P}(\|\mathbb{D}_j^{\mathrm{c}\prime} \mathbf{H}_n \boldsymbol{\theta}_n\|_2 > \lambda_{n2} r_n/8, \exists\, j \notin A_{**}) \\ &\quad + \mathbb{P}(\|\lambda_{n2} n^{-1} \mathbb{D}_j^{\mathrm{c}\prime} \mathbb{D}_{A_{**}}^{\mathrm{c}} \boldsymbol{\Omega}_{A_*}^{-1} \mathbf{v}_n\|_2 > \lambda_{n2} r_n/8, \exists\, j \notin A_{**}) \end{aligned} \tag{33}$$

so we find upper bounds of the four terms in inequality (33). For the first term, we have

$$\begin{aligned}\max_{j\notin A_{**}} \|\mathbb{D}_j^{\mathrm{c}\prime} \mathbb{D}_{A_* \cap A_{**}^{\mathrm{c}}}^{\mathrm{c}} \boldsymbol{\beta}_{A_* \cap A_{**}^{\mathrm{c}}}\|_2 &\leqslant n \max_{j\notin A_{**}} \|n^{-1/2} \mathbb{D}_j^{\mathrm{c}\prime}\|_2 \|n^{-1/2} \mathbb{D}_{A_* \cap A_{**}^{\mathrm{c}}}^{\mathrm{c}}\|_2 \|\boldsymbol{\beta}_{A_* \cap A_{**}^{\mathrm{c}}}\|_2 \\ &= \mathbf{O}_p\{n\, \rho_{\max}^{1/2}(\boldsymbol{\Omega}_{A_{**}^{\mathrm{c}}})\, \rho_{\max}^{1/2}(\boldsymbol{\Omega}_{A_{**}}) m_n^{1/2}\} = \mathbf{O}_p(n m_n^{-1/2}). \end{aligned}$$

Then, for some generic constant C,

$$\begin{aligned}\mathbb{P}\left(\|\mathbb{D}_j^{\mathrm{c}\prime} \mathbb{D}_{A_* \cap A_{**}^{\mathrm{c}}}^{\mathrm{c}} \boldsymbol{\beta}_{A_* \cap A_{**}^{\mathrm{c}}}\|_2 > \frac{\lambda_{n2} r_n}{8}, \exists\, j \notin A_{**}\right) &\leqslant \mathbb{P}\left(\max_{j\notin A_{**}} \|\mathbb{D}_j^{\mathrm{c}\prime} \mathbb{D}_{A_* \cap A_{**}^{\mathrm{c}}}^{\mathrm{c}} \boldsymbol{\beta}_{A_* \cap A_{**}^{\mathrm{c}}}\|_2 > \frac{C\lambda_{n2} r_n}{8}\right) \\ &\leqslant \mathbb{P}\left(n m_n^{-1/2} > \frac{C\lambda_{n2} r_n}{8}\right) = \mathbb{P}\left(\frac{n m_n^{-1/2}}{\lambda_{n2} r_n} > \frac{C}{8}\right) \to 0, \end{aligned} \tag{34}$$

where result (34) holds by assumption 2, part (b). For the second term, let $s_n = J - |A_{**}|$. Since $\rho_{\max}(\mathbf{H}_n) = \rho_{\max}(\mathbb{I} - P_{A_{**}}) = 1 - \rho_{\min}(P_{A_{**}})$ and $P_{A_{**}}$ is a non-negative definite matrix, $\rho_{\max}(\mathbf{H}_n) \leqslant 1$. By lemma 3 with

$\mathbf{M}_n = \mathbf{L}\mathbf{H}_n$, and using the fact that $\rho_{\max}(\mathbf{L}\mathbf{H}_n) \leqslant \rho_{\max}(\mathbf{L})\rho_{\max}(\mathbf{H}_n)$, we have

$$\begin{aligned}\mathbb{E}\left(\max_{j\notin A_{**}} n^{-1/2}\|\mathbb{D}_j^{\mathrm{c}\prime}\mathbf{H}_n\boldsymbol{\epsilon}_n\|_2\right) &= \mathbb{E}\left(\max_{j\notin A_{**}} n^{-1/2}\sqrt{\sum_{l=1}^{m_n}|\mathbb{D}_{jl}^{\mathrm{c}\,\prime}\mathbf{H}_n\boldsymbol{\epsilon}_n|^2}\right)\\ &\leqslant \mathbb{E}\left\{\max_{\substack{j\notin A_{**}\\ 1\leqslant l\leqslant m_n}}\left(\frac{m_n}{n}\right)^{1/2}|\mathbf{a}_{jl}'\mathbf{L}\mathbf{H}_n\boldsymbol{\epsilon}_n|\right\} = \mathbf{O}[\surd\{\rho_{\max}^2(\mathbf{L})m_n\log(s_n m_n)\}]. \qquad (35)\end{aligned}$$

Thus, by Markov's inequality,

$$\begin{aligned}\mathbb{P}\left(\|\mathbb{D}_j^{\mathrm{c}\prime}\mathbf{H}_n\boldsymbol{\epsilon}_n\|_2 > \frac{\lambda_{n2}r_n}{8}, \exists\, j\notin A_{**}\right) &\leqslant \mathbb{P}\left(\max_{j\notin A_{**}} n^{-1/2}\|\mathbb{D}_j^{\mathrm{c}\prime}\mathbf{H}_n\boldsymbol{\epsilon}_n\|_2 > \frac{Cn^{-1/2}\lambda_{n2}r_n}{8}\right)\\ &\leqslant \mathbf{O}\left[\frac{\surd\{\rho_{\max}^2(\mathbf{L})m_n\log(s_n m_n)\}}{C\lambda_{n2}r_n}\right] \to 0, \qquad (36)\end{aligned}$$

where C is a generic constant and result (36) holds by assumption 2, part (b). For the third term,

$$\max_{j\notin A_{**}}\|\mathbb{D}_j^{\mathrm{c}\prime}\mathbf{H}_n\boldsymbol{\theta}_n\|_2 \leqslant n^{1/2}\max_{j\notin A_{**}}\|n^{-1}\mathbb{D}_j^{\mathrm{c}\prime}\mathbb{D}_j^{\mathrm{c}}\|_2^{1/2}\|\mathbf{H}_n\|_2\|\boldsymbol{\theta}_n\|_2 = \mathbf{O}\{n\,\rho_{\max}^{1/2}(\boldsymbol{\Omega}_{A_{**}^{\mathrm{c}}})m_n^{-\tau}\} = \mathbf{O}(nm_n^{-\tau-1/2}).$$

Therefore, for some generic constant C,

$$\begin{aligned}\mathbb{P}\left(\|\mathbb{D}_j^{\mathrm{c}\prime}\mathbf{H}_n\boldsymbol{\theta}_n\|_2 > \frac{\lambda_{n2}r_n}{6}, \exists\, j\notin A_{**}\right) &\leqslant \mathbb{P}\left(\max_{j\notin A_{**}}\|\mathbb{D}_j^{\mathrm{c}\prime}\mathbf{H}_n\boldsymbol{\theta}_n\|_2 > \frac{C\lambda_{n2}r_n}{8}\right)\\ &\leqslant \mathbb{P}\left(nm_n^{-\tau-1/2} > \frac{C\lambda_{n2}r_n}{8}\right) = \mathbb{P}\left(\frac{n}{\lambda_{n2}r_n m_n^{(2\tau+1)/2}} > \frac{C}{8}\right) \to 0, \qquad (37)\end{aligned}$$

where result (37) is implied by assumption 2, part (b). Finally, using equation (28) we have

$$\begin{aligned}\max_{j\notin A_{**}}\|\lambda_{n2}n^{-1}\mathbb{D}_j^{\mathrm{c}\prime}\mathbb{D}_{A_{**}}^{\mathrm{c}}\boldsymbol{\Omega}_{A_{**}}^{-1}\mathbf{v}_n\|_2 &\leqslant \lambda_{n2}\max_{j\notin A_{**}}\|n^{-1/2}\mathbb{D}_j^{\mathrm{c}\prime}\|_2\|n^{-1/2}\mathbb{D}_{A_{**}}^{\mathrm{c}}\boldsymbol{\Omega}_{A_{**}}^{-1/2}\|_2\|\boldsymbol{\Omega}_{A_{**}}^{-1/2}\|_2\|\mathbf{v}_n\|_2\\ &= \mathbf{O}_p\{\lambda_{n2}\,\rho_{\max}^{1/2}(\boldsymbol{\Omega}_{A_{**}^{\mathrm{c}}})\rho_{\min}^{-1/2}(\boldsymbol{\Omega}_{A_{**}})k_n\} = \mathbf{O}_p\{\lambda_{n2}(m_n^{-1}r_n^{-1/2} + m_n^{-1/2})\}.\end{aligned}$$

Then, for some generic constant C,

$$\begin{aligned}\mathbb{P}\left(\|\lambda_{n2}n^{-1}\mathbb{D}_j^{\mathrm{c}\prime}\mathbb{D}_{A_{**}}^{\mathrm{c}}\boldsymbol{\Omega}_{A_{**}}^{-1}\mathbf{v}_n\|_2 > \frac{\lambda_{n2}r_n}{8}, \exists\, j\notin A_{**}\right) &\leqslant \mathbb{P}\left(\max_{j\notin A_{**}}\|\lambda_{n2}n^{-1}\mathbb{D}_j^{\mathrm{c}\prime}\mathbb{D}_{A_{**}}^{\mathrm{c}}\boldsymbol{\Omega}_{A_{**}}^{-1}\mathbf{v}_n\|_2 > \frac{C\lambda_{n2}r_n}{8}\right)\\ &\leqslant \mathbb{P}\left\{\lambda_{n2}(m_n^{-1}r_n^{-1/2} + m_n^{-1/2}) > \frac{C\lambda_{n2}r_n}{8}\right\}\\ &= \mathbb{P}\left(\frac{m_n^{-1}r_n^{-1/2} + m_n^{-1/2}}{r_n} > \frac{C}{8}\right) \to 0, \qquad (38)\end{aligned}$$

where result (38) holds since $r_n, m_n \to \infty$. Hence, by combining results (36), (37) and (38), result (30) follows.

A.3.2. Part (b)

Denote $\eta_* = \max_{j\in A_*} 1/\|\boldsymbol{\beta}_j\|_2$. Let $A_{***} = A_* \cup \{j : \|\hat{\boldsymbol{\beta}}_{\mathrm{AGL},j}\|_2 > 0\}$. Note that $J_0 = |A_{***}|$. Define $\boldsymbol{\vartheta}_{A_{***}} = \mathbf{Z}^{\mathrm{c}} - \mathbb{D}_{A_{***}}^{\mathrm{c}}\boldsymbol{\beta}_{A_{***}}$ and denote $\boldsymbol{\vartheta}_{A_{***}}^*$ and $\boldsymbol{\epsilon}_{A_{***}}^*$ as the projections of $\boldsymbol{\vartheta}_{A_{***}}$ and $\boldsymbol{\epsilon}_n$ to the span of $\mathbb{D}_{A_{***}}^{\mathrm{c}}$. Then, in a similar way to that in part (b) of theorem 1, we can show that

$$\begin{aligned}\|\boldsymbol{\vartheta}_{A_{***}}^*\|_2^2 &\leqslant \|\boldsymbol{\epsilon}_{A_{***}}^*\|_2^2 + \mathbf{O}_p(1 + |A_{***}|nm_n^{-2\tau})\\ &= \|(\mathbb{D}_{A_{***}}^{\mathrm{c}\prime}\mathbb{D}_{A_{***}}^{\mathrm{c}})^{-1/2}\mathbb{D}_{A_{***}}^{\mathrm{c}\prime}\boldsymbol{\epsilon}_n\|_2^2 + \mathbf{O}_p(1 + |A_{***}|nm_n^{-2\tau})\\ &= \mathbf{O}_{\mathrm{p}}\left\{\frac{|A_{***}|n^{\alpha}\rho_{\max}^2(\mathbf{L})m_n\log(J_0 m_n)}{c_1 h_n} + 1 + |A_{***}|nm_n^{-2\tau}\right\}. \qquad (39)\end{aligned}$$

In a similar way to obtain expression (16), we can also show that

$$\|\hat{\boldsymbol{\beta}}_{\mathrm{AGL},A_{***}} - \boldsymbol{\beta}_{A_{***}}\|_2^2 \leqslant \frac{8\|\boldsymbol{\vartheta}^*_{A_{***}}\|_2^2}{nc_1h_n} + \frac{4\lambda_{n2}^2|A_{***}|\eta_*}{n^2c_1^2h_n^2}; \tag{40}$$

thus, by expressions (39) and (40), we obtain part (b) of theorem 3.

A.4. Proof of theorem 4

The proof is similar to that for theorem 2.

References

Antoniadis, A. and Fan, J. (2001) Regularization of wavelet approximation (with discussion). *J. Am. Statist. Ass.*, **96**, 939–967.

Belloni, A. and Chernozhukov, V. (2013) Least squares after model selection in high-dimensional sparse models. *Bernoulli*, **19**, 521–547.

Chu, T., Zhu, J. and Wang, H. (2011) Penalized maximum likelihood estimation and variable selection in geostatistics. *Ann. Statist.*, **39**, 2607–2625.

Cressie, N. (2015) *Statistics for Spatial Data*, revised edn. Hoboken: Wiley.

Cressie, N. and Chan, N. H. (1989) Spatial modeling of regional variables. *J. Am. Statist. Ass.*, **84**, 393–401.

Fan, Y. and Tang, C. Y. (2013) Tuning parameter selection in high dimensional penalized likelihood. *J. R. Statist. Soc.* B, **75**, 531–552.

Fu, R., Thurman, A. L., Chu, T., Steen-Adams, M. M. and Zhu, J. (2013) On estimation and selection of autologistic regression models via penalized pseudolikelihood. *J. Agric. Biol. Environ. Statist.*, **18**, 429–449.

Gupta, S. (2012) A note on the asymptotic distribution of LASSO estimator for correlated data. *Sankhya* A, **74**, 10–28.

Haustein, K. O. (2006) Smoking and poverty. *Eur. J. Prevn. Cardiol.*, **13**, 312–318.

Hoeting, J. A., Davis, R. A., Merton, A. A. and Thompson, S. E. (2006) Model selection for geostatistical models. *Ecol. Appl.*, **16**, 87–98.

Horn, R. A. and Johnson, C. A. (1985) *Matrix Analysis*. Cambridge: Cambridge University Press.

Hsu, N.-J., Hung, H.-L. and Chang, Y.-M. (2008) Subset selection for vector autoregressive processes using Lasso. *Computnl Statist. Data Anal.*, **52**, 3645–3657.

Huang, H. C. and Chen, C. S. (2007) Optimal geostatistical model selection. *J. Am. Statist. Ass.*, **102**, 1009–1024.

Huang, J., Horowitz, J. L. and Wei, F. (2010a) Variable selection in nonparametric additive models. *Ann. Statist.*, **38**, 2282–2313.

Huang, H. C., Hsu, N. J., Theobald, D. M. and Breidt, F. J. (2010b) Spatial Lasso with applications to GIS model selection. *J. Computnl Graph. Statist.*, **19**, 963–983.

Kneib, T., Hothorn, T. and Tutz, G. (2009) Variable selection and model choice in geoadditive regression models. *Biometrics*, **65**, 626–634.

Lai, R. C. S., Huang, H. and Lee, T. C. M. (2012) Fixed and random effects selection in nonparametric additive mixed models. *Electron. J. Statist.*, **6**, 810–842.

Lin, Y. and Zhang, H. (2006) Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist.*, **34**, 2272–2297.

Liu, H. and Yu, B. (2013) Asymptotic properties of Lasso+mLS and Lasso+Ridge in sparse high-dimensional linear regression. *Electron. J. Statist.*, **7**, 3124–3169.

Meier, L., Van De Geer, S. and Bühlmann, P. (2009) High-dimensional additive modeling. *Ann. Statist.*, **37**, 3779–3821.

Nardi, Y. and Rinaldo, A. (2011) Autoregressive process modeling via the Lasso procedure. *J. Multiv. Anal.*, **102**, 528–549.

Nishii, R. (1984) Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.*, **12**, 758–765.

Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2009) Sparse additive models. *J. R. Statist. Soc.* B, **71**, 1009–1030.

Reyes, P. E., Zhu, J. and Aukema, B. H. (2012) Selection of spatial-temporal lattice models: assessing the impact of climate conditions on a mountain pine beetle outbreak. *J. Agric. Biol. Environ. Statist.*, **17**, 508–525.

Schumaker, L. (2007) *Spline Functions: Basic Theory*. Cambridge: Cambridge University Press.

Stein, M. L. (1999) *Interpolation of Spatial Data*. New York: Springer.

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. (2007) Regression coefficient and autoregressive order shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **69**, 63–78.

Wang, H. and Zhu, J. (2009) Variable selection in spatial regression via penalized least squares. *Can. J. Statist.*, **37**, 607–624.

Wendland, H. (2005) *Scattered Data Approximation*. Cambridge: Cambridge University Press.

Xu, G., Xiang, Y., Wang, S. and Lin, Z. (2012) Regularization and variable selection for infinite variance autoregressive models. *J. Statist. Planng Inf.*, **142**, 2545–2553.

Yuan, M. and Lin, Y. (2006) Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc.* B, **68**, 49–67.

Zhang, C. and Huang, J. (2008) The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Ann. Statist.*, **36**, 1567–1594.

Zhu, J., Huang, H.-C. and Reyes, P. E. (2010) On selection of spatial linear models for lattice data. *J. R. Statist. Soc.* B, **72**, 389–402.

Supporting information

Additional ‘supporting information’ may be found in the on-line version of this article:

‘Supplementary material to “Additive model building for spatial regression”’.