The Annals of Statistics
2011, Vol. 39, No. 5, 2607–2625
DOI: 10.1214/11-AOS919

PENALIZED MAXIMUM LIKELIHOOD ESTIMATION AND VARIABLE SELECTION IN GEOSTATISTICS

BY TINGJIN CHU, JUN ZHU[1] AND HAONAN WANG[2]

Colorado State University, University of Wisconsin, Madison, and Colorado State University

We consider the problem of selecting covariates in spatial linear models with Gaussian process errors. Penalized maximum likelihood estimation (PMLE) that enables simultaneous variable selection and parameter estimation is developed and, for ease of computation, PMLE is approximated by one-step sparse estimation (OSE). To further improve computational efficiency, particularly with large sample sizes, we propose penalized maximum covariance-tapered likelihood estimation (PMLE_T) and its one-step sparse estimation (OSE_T). General forms of penalty functions with an emphasis on smoothly clipped absolute deviation are used for penalized maximum likelihood. Theoretical properties of PMLE and OSE, as well as their approximations PMLE_T and OSE_T using covariance tapering, are derived, including consistency, sparsity, asymptotic normality and the oracle properties. For covariance tapering, a by-product of our theoretical results is consistency and asymptotic normality of maximum covariance-tapered likelihood estimates. Finite-sample properties of the proposed methods are demonstrated in a simulation study and, for illustration, the methods are applied to analyze two real data sets.

1. Introduction. Geostatistical models are popular tools for the analysis of spatial data in many disciplines. It is often of interest to estimate model parameters based on data at sampled locations and perform spatial interpolation (also known as Kriging) of a response variable at unsampled locations within a spatial domain of interest [2, 16, 17]. In addition, a practical issue that often arises is how to select the best model or a best subset of models among many competing ones [10]. Here we focus on selecting covariates in a spatial linear model, which we believe is a problem that is underdeveloped in both theory and methodology despite its importance in geostatistics. The spatial linear model for a response variable under consideration has two additive components: a fixed linear regression term and a stochastic error term. We assume that the error term follows a Gaussian process

Received July 2010; revised August 2011.

[1]Supported in part by a USDA CSREES Hatch project.

[2]Supported in part by NSF Grants DMS-07-06761, DMS-08-54903 and DMS-11-06975, and by the Air Force Office of Scientific Research under contract number FA9550-10-1-0241.

MSC2010 subject classifications. Primary 62F12; secondary 62M30.

Key words and phrases. Covariance tapering, Gaussian process, model selection, one-step sparse estimation, SCAD, spatial linear model.

with mean zero and a covariance function that accounts for spatial dependence. Our chief objective is to develop a set of new methods for the selection of covariates and establish their asymptotic properties. Moreover, we devise efficient algorithms for computation, making these methods feasible for practical usage.

For linear regression with independent errors, variable selection has been widely studied in the literature. The more traditional methods often involve hypothesis testing such as F-tests in a stepwise selection procedure [3]. An alternative approach is to select models using information discrepancy such as a Kolmogorov–Smirnov, Kullback–Leibler or Hellinger discrepancy [13]. In recent years, penalized methods are becoming increasingly popular for variable selection. For example, Tibshirani [18] developed a least absolute shrinkage and selection operator (LASSO), whereas Fan and Li [7] proposed a nonconcave penalized likelihood method with a smoothly clipped absolute deviation (SCAD) penalty. Efron et al. [5] devised least angle regression (LARS) algorithms, which allow computing all LASSO estimates along a path of its tuning parameters at a low computational order. More recently, Zou [23] improved LASSO and the resulting adaptive LASSO enjoys the oracle properties as SCAD, in terms of selecting the true model. Zou and Li [24] proposed one-step sparse estimation in the nonconcave penalized likelihood approach, which retains the oracle properties and utilizes LARS algorithms.

For spatial linear models in geostatistics, in contrast, statistical methods for a principled selection of covariates are limited. Hoeting et al. [10] suggested Akaike's information criterion (AIC) with a finite-sample correction for variable selection. Like information-based selection in general, computation can be costly especially when the number of covariates and/or the sample sizes are large. Thus, these authors considered only a subset of the covariates that may be related to the abundance of the orange-throated whiptail lizard in southern California, in order to make it tractable to evaluate their AIC-based model selection. Huang and Chen [11] developed a model selection criterion in geostatistics, but for the purpose of Kriging rather than selection of covariates. Further, Wang and Zhu [21] proposed penalized least squares (PLS) for a spatial linear model where the error process is assumed to be strong mixing without the assumption of a Gaussian process. This method includes spatial autocorrelation only indirectly in the sense that the objective function involves a sum of squared errors ignoring spatial dependence. A spatial block bootstrap is then used to account for spatial dependence when estimating the variance of PLS estimates.

Here we take an alternative, parametric approach and assume that the errors in the spatial linear model follow a Gaussian process. Our main innovation here is to incorporate spatial dependence directly into a penalized likelihood function and achieve greater efficiency in the resulting penalized maximum likelihood estimates (PMLE). Unlike computation of PLS estimates which is on the same order as ordinary least squares estimates, however, penalized likelihood function for a spatial linear model will involve operations of a covariance matrix of the same size as the number of observations. Thus, the computational cost can be prohibitively high as

the sample size becomes large. It is essential that our new methods address this issue. To that end, we utilize one-step sparse estimation (OSE) and LARS algorithms in the computation of PMLE to gain computational efficiency. In addition, we explore covariance tapering, which further reduces computational cost by replacing the exact covariance matrix with a sparse one [4, 9, 12]. We establish the asymptotic properties of both PMLE and OSE, as well as their covariance-tapered counterparts. As a by-product, we establish new results for covariance-tapered MLE which, to the best of our knowledge, have not been established before and can be of independent interest.

The remainder of the paper is organized as follows. In Section 2 we develop penalized maximum covariance-tapered likelihood estimation (PMLE_T) that enables simultaneous variable selection and parameter estimation, as well as an approximation of the PMLE_T by one-step sparse estimation (OSE_T) to enhance computational efficiency. PMLE and OSE are regarded as a special case of PMLE_T and OSE_T. We establish asymptotic properties of PMLE and OSE in Section 3 and those of PMLE_T and OSE_T under covariance tapering in Section 4. In Section 5 finite-sample properties of the proposed methods are investigated in a simulation study and, for illustration, the methods are applied to analyze two real data sets. We outline the technical proofs in Appendices A.1 and A.2.

2. Maximum likelihood estimation: Penalization and covariance tapering.

2.1. *Spatial linear model and maximum likelihood estimation.* For a spatial domain of interest R in $\mathbb{R}^d$, we consider a spatial process $\{y(\mathbf{s}) : \mathbf{s} \in R\}$ such that

$$y(\mathbf{s}) = \mathbf{x}(\mathbf{s})^T \boldsymbol{\beta} + \varepsilon(\mathbf{s}), \tag{2.1}$$

where $\mathbf{x}(\mathbf{s}) = (x_1(\mathbf{s}), \ldots, x_p(\mathbf{s}))^T$ is a $p \times 1$ vector of covariates at location $\mathbf{s}$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a $p \times 1$ vector of regression coefficients. We assume that the error process $\{\varepsilon(\mathbf{s}) : \mathbf{s} \in R\}$ is a Gaussian process with mean zero and a covariance function

$$\gamma(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta}) = \operatorname{cov}\{\varepsilon(\mathbf{s}), \varepsilon(\mathbf{s}')\}, \tag{2.2}$$

where $\mathbf{s}, \mathbf{s}' \in R$ and $\boldsymbol{\theta}$ is a $q \times 1$ vector of covariance function parameters.

Let $\mathbf{s}_1, \ldots, \mathbf{s}_N$ denote N sampling sites in R. Let $\mathbf{y} = (y(\mathbf{s}_1), \ldots, y(\mathbf{s}_N))^T$ denote an $N \times 1$ vector of response variables and $\mathbf{x}_j = (x_j(\mathbf{s}_1), \ldots, x_j(\mathbf{s}_N))^T$ denote an $N \times 1$ vector of the jth covariate with $j = 1, \ldots, p$, at the N sampling sites. Further, let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_p]$ denote an $N \times p$ design matrix of covariates and $\boldsymbol{\Gamma} = [\gamma(\mathbf{s}_i, \mathbf{s}_{i'}; \boldsymbol{\theta})]_{i,i'=1}^N$ denote an $N \times N$ covariance matrix. In this paper, we consider general forms for the the covariance matrix $\boldsymbol{\Gamma}$ and describe suitable regularity conditions in Sections 3 and 4. By (2.1) and (2.2), we have

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Gamma}). \tag{2.3}$$

Let $\boldsymbol{\eta} = (\boldsymbol{\beta}^T, \boldsymbol{\theta}^T)^T$ denote a $(p+q) \times 1$ vector of model parameters consisting of both regression coefficients $\boldsymbol{\beta}$ and covariance function parameters $\boldsymbol{\theta}$. By (2.3), the log-likelihood function of $\boldsymbol{\eta}$ is

$$\begin{aligned}\ell(\boldsymbol{\eta}; \mathbf{y}, \mathbf{X}) = &-(N/2)\log(2\pi) - (1/2)\log|\boldsymbol{\Gamma}| \\ &- (1/2)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \boldsymbol{\Gamma}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).\end{aligned} \tag{2.4}$$

Let $\widehat{\boldsymbol{\eta}}_{\mathrm{MLE}} = \arg\max_{\boldsymbol{\eta}}\{\ell(\boldsymbol{\eta}; \mathbf{y}, \mathbf{X})\}$ denote the maximum likelihood estimate (MLE) of $\boldsymbol{\eta}$.

2.2. *Covariance tapering and penalized maximum likelihood.* It is well known that computation of MLE for a spatial linear model is of order N^3 and can be very demanding when the sample size N increases [2]. There are various approaches to alleviating the computational cost. Here we consider covariance tapering, which could effectively reduce our computational cost in practice. Furrer et al. [9] considered tapering for Kriging and demonstrated that not only tapering enhances computational efficiency but also achieves asymptotically optimality in terms of mean squared prediction errors under infill asymptotics. For parameter estimation via maximum likelihood, Kaufman et al. [12] established consistency of tapered MLE, whereas Du et al. [4] established the asymptotic distribution, also under infill asymptotics. However, both Kaufman et al. [12] and Du et al. [4] focused on the parameters in the Matérn family of covariance functions and did not consider estimation of the regression coefficients. In contrast, our primary interest is in the estimation of regression coefficients and we investigate the asymptotic properties under increasing domain asymptotics, which, to the best of our knowledge, have not been established in the literature before.

Recall that $\boldsymbol{\Gamma} = [\gamma(\mathbf{s}_i, \mathbf{s}_{i'})]_{i,i'=1}^N$ is the covariance matrix of $\mathbf{y}$. Assuming second-order stationarity and isotropy, we let $\gamma(d) = \gamma(\mathbf{s}, \mathbf{s}')$, where $d = \|\mathbf{s} - \mathbf{s}'\|$ is the lag distance between two sampling sites $\mathbf{s}$ and $\mathbf{s}'$ in R. Let $K_{\mathrm{T}}(d, \omega)$ denote a tapering function, which is an isotropic autocorrelation function when $0 < d < \omega$ and 0 when $d \geq \omega$, for a given threshold distance $\omega > 0$. Compactly supported correlation functions can be used as the tapering functions [22]. For example,

$$K_{\mathrm{T}}(d, \omega) = (1 - d/\omega)_+, \tag{2.5}$$

where $x_+ = \max\{x, 0\}$, in which case the correlation is 0 at lag distance greater than the threshold distance ω. Let $\boldsymbol{\Delta}(\omega) = [K_{\mathrm{T}}(d_{ii'}, \omega)]_{i,i'=1}^N$ denote an $N \times N$ tapering matrix. Then a tapered covariance matrix of $\boldsymbol{\Gamma}$ is defined as $\boldsymbol{\Gamma}_{\mathrm{T}} = \boldsymbol{\Gamma} \circ \boldsymbol{\Delta}(\omega)$, where $\circ$ denotes the Hadamard product (i.e., elementwise product).

We approximate the log-likelihood function by replacing $\boldsymbol{\Gamma}$ in (2.4) with the tapered covariance matrix $\boldsymbol{\Gamma}_{\mathrm{T}}$ and obtain a covariance-tapered log-likelihood function

$$\begin{aligned}\ell_{\mathrm{T}}(\boldsymbol{\eta}; \mathbf{y}, \mathbf{X}) = &-(N/2)\log(2\pi) - (1/2)\log|\boldsymbol{\Gamma}_{\mathrm{T}}| \\ &- (1/2)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \boldsymbol{\Gamma}_{\mathrm{T}}^{-1}(\mathbf{y} - \boldsymbol{X\beta}).\end{aligned} \tag{2.6}$$

We let $\widehat{\boldsymbol{\eta}}_{\mathrm{MLE_T}} = \arg\max_{\boldsymbol{\eta}}\{\ell_{\mathrm{T}}(\boldsymbol{\eta}; \mathbf{y}, \mathbf{X})\}$ denote the maximum covariance-tapered likelihood estimate ($\mathrm{MLE_T}$) of $\boldsymbol{\eta}$.

Let $\boldsymbol{\Gamma}_{k,\mathrm{T}} = \partial\boldsymbol{\Gamma}_{\mathrm{T}}/\partial\theta_k = \boldsymbol{\Gamma}_k \circ \boldsymbol{\Delta}(\omega)$, $\boldsymbol{\Gamma}_{\mathrm{T}}^k = \partial\boldsymbol{\Gamma}_{\mathrm{T}}^{-1}/\partial\theta_k = \boldsymbol{\Gamma}^k \circ \boldsymbol{\Delta}(\omega)$, $\boldsymbol{\Gamma}_{kk',\mathrm{T}} = \partial^2\boldsymbol{\Gamma}_{\mathrm{T}}/\partial\theta_k\,\partial\theta_{k'} = \boldsymbol{\Gamma}_{kk'} \circ \boldsymbol{\Delta}(\omega)$, and $\boldsymbol{\Gamma}_{\mathrm{T}}^{kk'} = \partial^2\boldsymbol{\Gamma}_{\mathrm{T}}^{-1}/\partial\theta_k\,\partial\theta_{k'} = \boldsymbol{\Gamma}^{kk'} \circ \boldsymbol{\Delta}(\omega)$ denote the covariance-tapered version of $\boldsymbol{\Gamma}_k$, $\boldsymbol{\Gamma}^k$, $\boldsymbol{\Gamma}_{kk'}$ and $\boldsymbol{\Gamma}^{kk'}$, respectively. From (2.6), $\ell'_{\mathrm{T}}(\boldsymbol{\beta}) = \mathbf{X}^T\boldsymbol{\Gamma}_{\mathrm{T}}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ and the kth element of $\ell'_{\mathrm{T}}(\boldsymbol{\theta})$ is $-(1/2)\operatorname{tr}(\boldsymbol{\Gamma}_{\mathrm{T}}^{-1}\boldsymbol{\Gamma}_{k,\mathrm{T}}) - (1/2)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\boldsymbol{\Gamma}_{\mathrm{T}}^k(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$. Moreover, $\ell''_{\mathrm{T}}(\boldsymbol{\beta}, \boldsymbol{\beta}) = -\mathbf{X}^T\boldsymbol{\Gamma}_{\mathrm{T}}^{-1}\mathbf{X}$, the kth column of $\ell''_{\mathrm{T}}(\boldsymbol{\beta}, \boldsymbol{\theta})$ is $\mathbf{X}^T\boldsymbol{\Gamma}_{\mathrm{T}}^k(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$, and the (k, k')th entry of $\ell''_{\mathrm{T}}(\boldsymbol{\theta}, \boldsymbol{\theta})$ is $-(1/2)\{\operatorname{tr}(\boldsymbol{\Gamma}_{\mathrm{T}}^{-1}\boldsymbol{\Gamma}_{kk',\mathrm{T}} + \boldsymbol{\Gamma}_{\mathrm{T}}^k\boldsymbol{\Gamma}_{k',\mathrm{T}}) + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\boldsymbol{\Gamma}_{\mathrm{T}}^{kk'}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\}$. Since $E\{-\ell''_{\mathrm{T}}(\boldsymbol{\beta}, \boldsymbol{\theta})\} = \mathbf{0}$, the covariance-tapered information matrix of $\boldsymbol{\eta}$ is $\mathbf{I}_{\mathrm{T}}(\boldsymbol{\eta}) = \operatorname{diag}\{\mathbf{I}_{\mathrm{T}}(\boldsymbol{\beta}), \mathbf{I}_{\mathrm{T}}(\boldsymbol{\theta})\}$, where $\mathbf{I}_{\mathrm{T}}(\boldsymbol{\beta}) = E\{-\ell''_{\mathrm{T}}(\boldsymbol{\beta}, \boldsymbol{\beta})\} = \mathbf{X}^T\boldsymbol{\Gamma}_{\mathrm{T}}^{-1}\mathbf{X}$ and the (k, k')th entry of $\mathbf{I}_{\mathrm{T}}(\boldsymbol{\theta}) = E\{-\ell''_{\mathrm{T}}(\boldsymbol{\theta}, \boldsymbol{\theta})\}$ is $t_{kk',\mathrm{T}}/2$ with $t_{kk',\mathrm{T}} = \operatorname{tr}(\boldsymbol{\Gamma}_{\mathrm{T}}^{-1}\boldsymbol{\Gamma}_{k,\mathrm{T}}\boldsymbol{\Gamma}_{\mathrm{T}}^{-1}\boldsymbol{\Gamma}_{k',\mathrm{T}}) = \operatorname{tr}(\boldsymbol{\Gamma}_{\mathrm{T}}\boldsymbol{\Gamma}_{\mathrm{T}}^k \times \boldsymbol{\Gamma}_{\mathrm{T}}\boldsymbol{\Gamma}_{\mathrm{T}}^{k'})$.

Now, we define a covariance-tapered penalized log-likelihood function as

$$Q_{\mathrm{T}}(\boldsymbol{\eta}) = \ell_{\mathrm{T}}(\boldsymbol{\eta}; \mathbf{y}, \mathbf{X}) - N\sum_{j=1}^{p} p_\lambda(|\beta_j|), \tag{2.7}$$

where $\ell_{\mathrm{T}}(\boldsymbol{\eta}; \mathbf{y}, \mathbf{X})$ is a covariance-tapered log-likelihood function as defined in (2.6). Moreover, we let $\widehat{\boldsymbol{\eta}}_{\mathrm{PMLE_T}} = \arg\max_{\boldsymbol{\eta}}\{Q_{\mathrm{T}}(\boldsymbol{\eta})\}$ denote the penalized maximum covariance-tapered likelihood estimate ($\mathrm{PMLE_T}$) of $\boldsymbol{\eta}$.

For penalty functions, we mainly consider smoothly clipped absolute deviation (SCAD) defined as

$$p_\lambda(\beta) = \begin{cases} \lambda|\beta|, & \text{if } |\beta| \le \lambda, \\ \lambda^2 + (a-1)^{-1}(a\lambda|\beta| - \beta^2/2 - a\lambda^2 + \lambda^2/2), & \\ & \text{if } \lambda < |\beta| \le a\lambda, \\ (a+1)\lambda^2/2, & \text{if } |\beta| > a\lambda, \end{cases} \tag{2.8}$$

for some $a > 2$ [6]. For i.i.d. error in standard linear regression, variable selection and parameter estimation under the SCAD penalty are shown to possess three desirable properties: unbiasedness, sparsity and continuity [7]. For spatial linear regression (2.1), these properties continue to hold for the SCAD penalty following arguments similar to those in Wang and Zhu [21].

To compute $\mathrm{PMLE_T}$ under the SCAD penalty, Fan and Li [7] proposed a locally quadratic approximation (LQA) of the penalty function and a Newton–Raphson algorithm. Although fast, a drawback of the LQA algorithm is that once a regression coefficient is shrunk to zero, it remains to be zero in the remainder iterations. More recently, Zou and Li [24] developed a unified algorithm to improve computational efficiency, which, unlike the LQA algorithm, is based on the locally linear approximation (LLA) of the penalty function. Moreover, Zou and Li [24] proposed one-step LLA estimation that approximates the solution after just one iteration in a Newton–Raphson-type algorithm starting at the MLE. We extend this one-step LLA estimation to approximate $\mathrm{PMLE_T}$ for the spatial linear model as follows.

ALGORITHM 1. At the initialization step, we let $\boldsymbol{\eta}_{\mathrm{T}}^{(0)} = \widehat{\boldsymbol{\eta}}_{\mathrm{MLE_T}}$ with $\boldsymbol{\beta}_{\mathrm{T}}^{(0)} = \widehat{\boldsymbol{\beta}}_{\mathrm{MLE_T}}$ and $\boldsymbol{\theta}_{\mathrm{T}}^{(0)} = \widehat{\boldsymbol{\theta}}_{\mathrm{MLE_T}}$. We then update $\boldsymbol{\beta}$ by maximizing

$$Q_{\mathrm{T}}^{*}(\boldsymbol{\beta}) = -(1/2)(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{T}\boldsymbol{\Gamma}_{\mathrm{T}}\big(\boldsymbol{\theta}_{\mathrm{T}}^{(0)}\big)^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) - N\sum_{j=1}^{p} p_{\lambda}'\big(|\beta_{j\mathrm{T}}^{(0)}|\big)|\beta_j| \tag{2.9}$$

with respect to $\boldsymbol{\beta}$, where the first term is from (2.6) and the second term is an LLA of the penalty function in (2.7). The resulting one-step sparse estimate (OSE) of $\boldsymbol{\beta}$ is denoted as $\widehat{\boldsymbol{\beta}}_{\mathrm{OSE_T}}$. We may also update $\boldsymbol{\theta}$ by maximizing (2.6) with respect to $\boldsymbol{\theta}$ given $\widehat{\boldsymbol{\beta}}_{\mathrm{OSE_T}}$. The resulting OSE of $\boldsymbol{\theta}$ is denoted as $\widehat{\boldsymbol{\theta}}_{\mathrm{OSE_T}}$. We let $\widehat{\boldsymbol{\eta}}_{\mathrm{OSE_T}} = (\widehat{\boldsymbol{\beta}}_{\mathrm{OSE_T}}^{T}, \widehat{\boldsymbol{\theta}}_{\mathrm{OSE_T}}^{T})^{T}$ denote the $\mathrm{OSE_T}$ of $\boldsymbol{\eta}$, which approximates $\widehat{\boldsymbol{\eta}}_{\mathrm{PMLE_T}}$.

It is worth mentioning an alternative covariance-tapered log-likelihood function [12],

$$\begin{aligned}\ell_{\mathrm{T2}}(\boldsymbol{\eta};\mathbf{y},\mathbf{X}) &= -(N/2)\log(2\pi) - (1/2)\log|\boldsymbol{\Gamma}_{\mathrm{T}}| \\ &\quad - (1/2)(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{T}\{\boldsymbol{\Gamma}_{\mathrm{T}}^{-1}\circ\boldsymbol{\Delta}(\omega)\}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}).\end{aligned} \tag{2.10}$$

If the alternative covariance tapering is used in Algorithm 1, the resulting estimates of parameters, especially the range parameter, tend to be more accurate, but require more time to compute $\boldsymbol{\Gamma}_{\mathrm{T}}^{-1}\circ\boldsymbol{\Delta}(\omega)$ than $\boldsymbol{\Gamma}_{\mathrm{T}}^{-1}$. For a numerical comparison, see Section 6.1 in Chu et al. [1].

Finally, two tuning parameters, λ and a, in the SCAD penalty (2.8) need to be estimated. For computational ease, we fix $a = 3.7$ as recommended by Fan and Li [7]. To determine λ, we use the Bayesian information criterion (BIC); see Wang et al. [20]. In particular, let

$$\widehat{\sigma}^{2}(\lambda) = N^{-1}\{\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}(\lambda)\}^{T}\boldsymbol{\Gamma}\{\widehat{\boldsymbol{\theta}}(\lambda)\}^{-1}\{\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}(\lambda)\}, \tag{2.11}$$

where $\widehat{\boldsymbol{\beta}}(\lambda)$ and $\widehat{\boldsymbol{\theta}}(\lambda)$ are the PMLE obtained for a given λ, and let

$$\mathrm{BIC}(\lambda) = N\log\{\widehat{\sigma}^{2}(\lambda)\} + k(\lambda)\log(N), \tag{2.12}$$

where $k(\lambda)$ is the number of nonzero regression coefficients [19]. Thus, an estimate of λ is $\widehat{\lambda} = \arg\min_{\lambda}\{\mathrm{BIC}(\lambda)\}$.

When $\boldsymbol{\Delta}(\omega)$ is a matrix of 1's, $\boldsymbol{\Gamma}_{\mathrm{T}} = \boldsymbol{\Gamma}$ and $\ell_{\mathrm{T}}(\cdot) = \ell(\cdot)$. Similarly, we henceforth obtain needed counterparts of the notation in this section under maximum likelihood without covariance tapering by omitting T. For details regarding such notation, see Section 2 of Chu et al. [1].

3. Asymptotic properties of PMLE and OSE.

3.1. *Notation and assumptions.* We let $\boldsymbol{\beta}_0 = (\beta_{10},\ldots,\beta_{p0})^{T} = (\boldsymbol{\beta}_{10}^{T}, \boldsymbol{\beta}_{20}^{T})^{T}$ denote the true regression coefficients, where without loss of generality $\boldsymbol{\beta}_{10}$ is an

$s \times 1$ vector of nonzero regression coefficients and $\boldsymbol{\beta}_{20} = \mathbf{0}$ is a $(p-s) \times 1$ zero vector. Let $\boldsymbol{\theta}_0$ denote the vector of true covariance function parameters.

We consider the asymptotic framework in Mardia and Marshall [14] and let n denote the stage of the asymptotics. In particular, write $R_n = R$, $N_n = N$, and $\lambda_n = \lambda$. Furthermore, define $a_n = \max_{1 \le j \le p}\{|p'_{\lambda_n}(|\beta_{j0}|)| : \beta_{j0} \neq 0\}$ and $b_n = \max_{1 \le j \le p}\{|p''_{\lambda_n}(|\beta_{j0}|)| : \beta_{j0} \neq 0\}$. Also, let $\boldsymbol{\phi}(\boldsymbol{\beta}) = (p'_\lambda(|\beta_1|)\operatorname{sgn}(\beta_1), \ldots, p'_\lambda(|\beta_p|)\operatorname{sgn}(\beta_p))^T$ and $\boldsymbol{\Phi}(\boldsymbol{\beta}) = \operatorname{diag}\{p''_\lambda(|\beta_1|), \ldots, p''_\lambda(|\beta_p|)\}$. Moreover, denote $\boldsymbol{\phi}_n(\boldsymbol{\beta}) = \boldsymbol{\phi}(\boldsymbol{\beta})$ and $\boldsymbol{\Phi}_n(\boldsymbol{\beta}) = \boldsymbol{\Phi}(\boldsymbol{\beta})$, both evaluated at λ_n. For all other quantities that depend on n, the stage n will be in either the left superscript or the right subscript.

Recall that ${}^n t_{kk'} = \operatorname{tr}({}^n\boldsymbol{\Gamma}^{-1\,n}\boldsymbol{\Gamma}_k\,{}^n\boldsymbol{\Gamma}^{-1\,n}\boldsymbol{\Gamma}_{k'})$. Let $\mu_1 \le \cdots \le \mu_{N_n}$ denote the eigenvalues of ${}^n\boldsymbol{\Gamma}$. For $l = 1, \ldots, N_n$, let μ_l^k denote the eigenvalues of ${}^n\boldsymbol{\Gamma}_k$ such that $|\mu_1^k| \le \cdots \le |\mu_{N_n}^k|$ and let $\mu_l^{kk'}$ denote the eigenvalues of ${}^n\boldsymbol{\Gamma}_{kk'}$ such that $|\mu_1^{kk'}| \le \cdots \le |\mu_{N_n}^{kk'}|$.

For an $N_n \times N_n$ matrix $\mathbf{A} = (a_{ij})_{i,j=1}^{N_n}$, the Frobenius, max and spectral norm are defined as $\|\mathbf{A}\|_F = (\sum_{i=1}^{N_n}\sum_{j=1}^{N_n} a_{ij}^2)^{1/2}$, $\|\mathbf{A}\|_{\max} = \max\{|a_{ij}| : i, j = 1, \ldots, N_n\}$ and $\|\mathbf{A}\|_s = \max\{|\mu_l(\mathbf{A})| : l = 1, \ldots, N_n\}$, where $\mu_l(\mathbf{A})$ is the lth eigenvalue of $\mathbf{A}$.

The following regularity conditions are assumed for Theorems 3.1 and 3.2:

(A.1) For $\boldsymbol{\theta} \in \Omega$ where Ω is an open subset of $\mathbb{R}^q$ such that $\boldsymbol{\eta} \in \mathbb{R}^p \times \Omega$, the covariance function $\gamma(\cdot, \cdot; \boldsymbol{\theta})$ is twice differentiable with respect to $\boldsymbol{\theta}$ with continuous second-order derivatives and is positive definite in the sense that, for any $N_n \ge 1$ and $\mathbf{s}_1, \ldots, \mathbf{s}_{N_n}$, the covariance matrix $\boldsymbol{\Gamma} = [\gamma(\mathbf{s}_i, \mathbf{s}_j; \boldsymbol{\theta})]_{i,j=1}^{N_n}$ is positive definite.

(A.2) There exist positive constants C, C_k and $C_{kk'}$, such that $\lim_{n\to\infty} \mu_{N_n} = C < \infty$, $\lim_{n\to\infty} |\mu_{N_n}^k| = C_k < \infty$, $\lim_{n\to\infty} |\mu_{N_n}^{kk'}| = C_{kk'} < \infty$ for all $k, k' = 1, \ldots, q$.

(A.3) For some $\delta > 0$, there exist positive constants D_k, $D_{kk'}$ and $D^*_{kk'}$ such that (i) $\|{}^n\boldsymbol{\Gamma}_k\|_F^{-2} = D_k N_n^{-1/2-\delta}$ for $k = 1, \ldots, q$; (ii) either $\|{}^n\boldsymbol{\Gamma}_k + {}^n\boldsymbol{\Gamma}_{k'}\|_F^{-2} = D_{kk'} N_n^{-1/2-\delta}$ or $\|{}^n\boldsymbol{\Gamma}_k - {}^n\boldsymbol{\Gamma}_{k'}\|_F^{-2} = D^*_{kk'} N_n^{-1/2-\delta}$ for any $k \neq k'$.

(A.4) For any $k, k' = 1, \ldots, q$, (i) ${}^n a_{kk'} = \lim_{n\to\infty}\{{}^n t_{kk'}({}^n t_{kk}\,{}^n t_{k'k'})^{-1/2}\}$ exists and $\mathbf{A}_n = ({}^n a_{kk'})_{k,k'=1}^q$ is nonsingular; (ii) $|{}^n t_{kk}\,{}^n t_{k'k'}{}^{-1}|$ and $|{}^n t_{k'k'}\,{}^n t_{kk}{}^{-1}|$ are bounded.

(A.5) The design matrix $\mathbf{X}$ has full rank p and is uniformly bounded in max norm with $\lim_{n\to\infty}(\mathbf{X}^T\mathbf{X})^{-1} = \mathbf{0}$.

(A.6) There exists a positive constant C_0, such that $\|{}^n\boldsymbol{\Gamma}^{-1}\|_s < C_0 < \infty$.

(A.7) For $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\boldsymbol{\theta} \in \Omega$, $N_n^{-1}\mathbf{I}_n(\boldsymbol{\beta}) \to \mathbf{J}(\boldsymbol{\beta})$ and $N_n^{-1}\mathbf{I}_n(\boldsymbol{\theta}) \to \mathbf{J}(\boldsymbol{\theta})$ as $n \to \infty$.

(A.8) $a_n = O(N_n^{-1/2})$ and $b_n \to 0$ as $n \to \infty$.

(A.9) There exist positive constants c_1 and c_2 such that, when $\beta_1, \beta_2 > c_1\lambda_n$, $|p''_{\lambda_n}(\beta_1) - p''_{\lambda_n}(\beta_2)| \le c_2|\beta_1 - \beta_2|$.

(A.10) $\lambda_n \to 0$, $N_n^{1/2}\lambda_n \to \infty$ as $n \to \infty$.
(A.11) $\liminf_{n\to\infty} \liminf_{\beta\to 0^+} \lambda_n^{-1} p'_{\lambda_n}(\beta) > 0$.

Conditions (A.2), (A.3)(i), (A.4)(i) and (A.5) are assumed in Mardia and Marshall [14]. Conditions (A.1) and (A.5) are standard assumptions for MLE, whereas (A.2), (A.3)(i), (A.4)(i) and (A.6) ensure smoothness, growth and convergence of the information matrix [14]. Together with (A.7), they yield a central limit theorem of $\ell'(\boldsymbol{\eta})$ and convergence in probability of $\ell''(\boldsymbol{\eta})$. For establishing Theorems 3.1 and 3.2, only the parts (i) of (A.3) and (A.4) are used. Moreover, the implicit asymptotic framework is increasing the domain, where the sample size N_n grows at the increase of the spatial domain R_n [14]. Finally, (A.8)–(A.11) are mild regularity conditions regarding the penalty function and are sufficient for Theorems 3.1 and 3.2 to hold [7] and [8].

3.2. *Consistency and asymptotic normality of PMLE.*

THEOREM 3.1. *Under* (A.1)–(A.9), *there exists*, *with probability tending to one*, *a local maximizer* ${}^n\widehat{\boldsymbol{\eta}}$ *of* $Q(\boldsymbol{\eta})$ *such that* $\|{}^n\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0\| = O_p(N_n^{-1/2} + a_n)$. *If*, *in addition*, (A.10)–(A.11) *hold*, *then* ${}^n\widehat{\boldsymbol{\eta}} = ({}^n\widehat{\boldsymbol{\beta}}_1^T, {}^n\widehat{\boldsymbol{\beta}}_2^T, {}^n\widehat{\boldsymbol{\theta}}^T)^T$ *satisfies*:

(i) *Sparsity*: ${}^n\widehat{\boldsymbol{\beta}}_2 = \mathbf{0}$ *with probability tending to* 1.
(ii) *Asymptotic normality*:

$$N_n^{1/2}\{\mathbf{J}(\boldsymbol{\beta}_{10}) + \boldsymbol{\Phi}_n(\boldsymbol{\beta}_{10})\}[{}^n\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + \{\mathbf{J}(\boldsymbol{\beta}_{10}) + \boldsymbol{\Phi}_n(\boldsymbol{\beta}_{10})\}^{-1}\boldsymbol{\phi}_n(\boldsymbol{\beta}_{10})] \xrightarrow{D} N(\mathbf{0}, \mathbf{J}(\boldsymbol{\beta}_{10})),$$

$$N_n^{1/2}({}^n\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{D} N(\mathbf{0}, \mathbf{J}(\boldsymbol{\theta}_0)^{-1}),$$

where $\mathbf{J}(\boldsymbol{\beta}_{10})$ *and* $\boldsymbol{\Phi}_n(\boldsymbol{\beta}_{10})$ *consist of the first* $s \times s$ *upper-left submatrix of* $\mathbf{J}(\boldsymbol{\beta}_0)$ *and* $\boldsymbol{\Phi}_n(\boldsymbol{\beta}_0)$, *respectively.*

Theorem 3.1 establishes the asymptotic properties of PMLE. Under (A.1)–(A.9), there exists a local maximizer converging to the true parameter at the rate $O_p(N_n^{-1/2} + a_n)$. Since $a_n = O(N_n^{-1/2})$ from (A.8), the local maximizer is root-N_n consistent. As shown in Fan and Li [7], the SCAD penalty function satisfies (A.8)–(A.11) by choosing an appropriate tuning parameter λ_n. Therefore, by Theorem 3.1, the PMLE under the SCAD penalty possesses the sparsity property and asymptotic normality. Moreover, when the sample size N_n is sufficiently large, $\boldsymbol{\Phi}_n(\boldsymbol{\beta}_{10})$ will be close to zero. That is, performance of the PMLE is asymptotically as efficient as the MLE of $\boldsymbol{\beta}_1$ when knowing $\boldsymbol{\beta}_2 = \mathbf{0}$. The arguments above hold for other penalty functions such as L_q penalty with $q < 1$, but not $q = 1$.

3.3. *Consistency and asymptotic normality of OSE.*

THEOREM 3.2. *Suppose that the initial value* ${}^n\boldsymbol{\eta}^{(0)}$ *satisfies* ${}^n\boldsymbol{\eta}^{(0)} - \boldsymbol{\eta}_0 = O_p(N_n^{-1/2})$. *For the SCAD penalty, under* (A.1)–(A.7) *and* (A.10), *the OSE* ${}^n\widehat{\boldsymbol{\eta}}_{\mathrm{OSE}} = ({}^n\widehat{\boldsymbol{\beta}}_{1,\mathrm{OSE}}^T, {}^n\widehat{\boldsymbol{\beta}}_{2,\mathrm{OSE}}^T, {}^n\widehat{\boldsymbol{\theta}}_{\mathrm{OSE}}^T)^T$ *satisfies*:

(i) *Sparsity*: ${}^n\widehat{\boldsymbol{\beta}}_{2,\mathrm{OSE}} = \mathbf{0}$ *with probability tending to* 1.
(ii) *Asymptotic normality*:

$$N_n^{1/2}({}^n\widehat{\boldsymbol{\beta}}_{1,\mathrm{OSE}} - \boldsymbol{\beta}_{10}) \xrightarrow{D} N(\mathbf{0}, \mathbf{J}(\boldsymbol{\beta}_{10})^{-1}),$$

$$N_n^{1/2}({}^n\widehat{\boldsymbol{\theta}}_{\mathrm{OSE}} - \boldsymbol{\theta}_0) \xrightarrow{D} N(\mathbf{0}, \mathbf{J}(\boldsymbol{\theta}_0)^{-1}),$$

where $\mathbf{J}(\boldsymbol{\beta}_{10})$ *consists of the first* $s \times s$ *upper-left submatrix of* $\mathbf{J}(\boldsymbol{\beta}_0)$.

Theorem 3.2 establishes the asymptotic properties of OSE such that the OSE is sparse and asymptotically normal under the SCAD penalty. The OSE for $\boldsymbol{\beta}_1$ and $\boldsymbol{\theta}$ has the same limiting distribution as the PMLE and thus achieves the same efficiency. In fact, Theorem 3.2 holds for another general class of penalty functions such that $p'_{\lambda_n}(\cdot) = \lambda_n p(\cdot)$ where $p'(\cdot)$ is continuous on $(0, \infty)$, and there is some $\alpha > 0$ such that $p'(\beta) = O(\beta^{-\alpha})$ as $\beta \to 0+$ [24]. Following similar arguments for the SCAD penalty in our Theorem 3.2 and those in Zou and Li [24], it can be shown that, if $N_n^{(1+\alpha)/2}\lambda_n \to \infty$ and $N_n^{1/2}\lambda_n \to 0$, Theorem 3.2 continues to hold. In practice, we set the initial value ${}^n\boldsymbol{\eta}^{(0)}$ to be the MLE ${}^n\widehat{\boldsymbol{\eta}}_{\mathrm{MLE}}$, as it satisfies the consistency condition.

4. Asymptotic properties under covariance tapering.

4.1. *Notation and assumptions.* In order to establish the asymptotic properties under covariance tapering, we continue to assume (A.1)–(A.11). We now restrict our attention to a second-order stationary error process in $\mathbb{R}^2$ with an isotropic covariance function $\gamma(d)$, where $d \geq 0$ is the lag distance. We also assume that the distance between any two sampling sites is greater than a constant [14]. As for the tapering function, we consider (2.5).

Let $\gamma_k(d) = \partial\gamma(d)/\partial\theta_k$, $\gamma_{kk'}(d) = \partial^2\gamma(d)/\partial\theta_k\,\partial\theta_{k'}$, for $k, k' = 1, \ldots, q$. Two additional regularity conditions are assumed for Theorems 4.2 and 4.3:

(A.12) $0 < \inf_n\{\omega_n N_n^{-1/2}\} \leq \sup_n\{\omega_n N_n^{-1/2}\} < \infty$, where $\omega_n = \omega$ is the threshold distance in the tapering function (2.5).

(A.13) There exists a nonincreasing function γ_0 with $\int_0^\infty u^2\gamma_0(u)\,du < \infty$ such that $\max\{|\gamma(u)|, |\gamma_k(u)|, |\gamma_{k,k'}(u)|\} \leq \gamma_0(u)$ for all $u \in (0, \infty)$ and $1 \leq k, k' \leq q$.

From (A.12), the threshold distance ω_n is bounded away from 0 and grows at the rate of $N_n^{1/2}$. The condition in (A.13) has to do with the covariance function. It can be shown that they hold for some of the commonly-used covariance functions such as the Matérn class. Details are given in Appendix D of Chu et al. [1].

4.2. *Consistency and asymptotic normality of PMLE*$_{\mathrm{T}}$.

PROPOSITION 4.1. *Under* (A.1)–(A.7) *and* (A.12)–(A.13), *the MLE*$_{\mathrm{T}}$ ${}^n\widehat{\boldsymbol{\eta}}_{\mathrm{MLE_T}}$ *is asymptotically normal with*

$$N_n^{1/2}({}^n\widehat{\boldsymbol{\eta}}_{\mathrm{MLE_T}} - \boldsymbol{\eta}_0) \xrightarrow{D} N(\mathbf{0}, \mathbf{J}(\boldsymbol{\eta}_0)^{-1}).$$

Proposition 4.1 establishes the asymptotic normality of MLE$_{\mathrm{T}}$. In particular, MLE and MLE$_{\mathrm{T}}$ have the same limiting distribution. This implies that, under the regularity conditions, covariance-tapered MLE achieves the same efficiency as MLE. Thus, in Algorithm 1 for computing the OSE$_{\mathrm{T}}$, we may set the initial parameter values to ${}^n\widehat{\boldsymbol{\eta}}_{\mathrm{MLE_T}}$.

THEOREM 4.2. *Under* (A.1)–(A.9) *and* (A.12)–(A.13), *there exists*, *with probability tending to one*, *a local maximizer* ${}^n\widehat{\boldsymbol{\eta}}_{\mathrm{T}}$ *of* $Q_{\mathrm{T}}(\boldsymbol{\eta})$ *defined in* (2.7) *such that* $\|{}^n\widehat{\boldsymbol{\eta}}_{\mathrm{T}} - \boldsymbol{\eta}_0\| = O_p(N_n^{-1/2} + a_n)$. *If*, *in addition*, (A.10)–(A.11) *hold*, *then* ${}^n\widehat{\boldsymbol{\eta}}_{\mathrm{T}} = ({}^n\widehat{\boldsymbol{\beta}}_{1,\mathrm{T}}^T, {}^n\widehat{\boldsymbol{\beta}}_{2,\mathrm{T}}^T, {}^n\widehat{\boldsymbol{\theta}}_{\mathrm{T}}^T)^T$ *satisfies*:

(i) *Sparsity*: ${}^n\widehat{\boldsymbol{\beta}}_{2,\mathrm{T}} = \mathbf{0}$ *with probability tending to* 1.

(ii) *Asymptotic normality*:

$$N_n^{1/2}\{\mathbf{J}(\boldsymbol{\beta}_{10}) + \boldsymbol{\Phi}_n(\boldsymbol{\beta}_{10})\}[{}^n\widehat{\boldsymbol{\beta}}_{1,\mathrm{T}} - \boldsymbol{\beta}_{10} + \{\mathbf{J}(\boldsymbol{\beta}_{10}) + \boldsymbol{\Phi}_n(\boldsymbol{\beta}_{10})\}^{-1}\boldsymbol{\phi}_n(\boldsymbol{\beta}_{10})]$$
$$\xrightarrow{D} N(\mathbf{0}, \mathbf{J}(\boldsymbol{\beta}_{10})),$$
$$N_n^{1/2}({}^n\widehat{\boldsymbol{\theta}}_{\mathrm{T}} - \boldsymbol{\theta}_0) \xrightarrow{D} N(\mathbf{0}, \mathbf{J}(\boldsymbol{\theta}_0)^{-1}),$$

where $\mathbf{J}(\boldsymbol{\beta}_{10})$ *and* $\boldsymbol{\Phi}_n(\boldsymbol{\beta}_{10})$ *consist of the first* $s \times s$ *upper-left submatrix of* $\mathbf{J}(\boldsymbol{\beta}_0)$ *and* $\boldsymbol{\Phi}_n(\boldsymbol{\beta}_0)$, *respectively.*

In Theorem 4.2, PMLE$_{\mathrm{T}}$ is shown to be consistent, sparse and asymptotically normal. In particular, PMLE$_{\mathrm{T}}$ has the same asymptotic distribution as PMLE in Theorem 3.1. That is, PMLE$_{\mathrm{T}}$ achieves the same efficiency and oracle property as PMLE asymptotically, yet in the mean time is more computationally efficient.

4.3. *Consistency and asymptotic normality of OSE*$_{\mathrm{T}}$.

THEOREM 4.3. *Suppose that the initial value* ${}^n\boldsymbol{\eta}_{\mathrm{T}}^{(0)}$ *in Algorithm* 1 *satisfies* ${}^n\boldsymbol{\eta}_{\mathrm{T}}^{(0)} - \boldsymbol{\eta}_0 = O_p(N_n^{-1/2})$. *For the SCAD penalty function*, *under* (A.1)–(A.7), (A.10) *and* (A.12)–(A.13), *the OSE*$_{\mathrm{T}}$ ${}^n\widehat{\boldsymbol{\eta}}_{\mathrm{OSE_T}} = ({}^n\widehat{\boldsymbol{\beta}}_{1,\mathrm{OSE_T}}^T, {}^n\widehat{\boldsymbol{\beta}}_{2,\mathrm{OSE_T}}^T, {}^n\widehat{\boldsymbol{\theta}}_{\mathrm{OSE_T}}^T)^T$ *satisfies*:

(i) *Sparsity*: ${}^n\widehat{\boldsymbol{\beta}}_{2,\mathrm{OSE_T}} = \mathbf{0}$ *with probability tending to* 1.

(ii) *Asymptotic normality*:

$$N_n^{1/2}({}^n\widehat{\boldsymbol{\beta}}_{1,\mathrm{OSE_T}} - \boldsymbol{\beta}_{10}) \xrightarrow{D} N(\mathbf{0}, \mathbf{J}(\boldsymbol{\beta}_{10})^{-1}),$$

$$N_n^{1/2}({}^n\widehat{\boldsymbol{\theta}}_{\mathrm{OSE_T}} - \boldsymbol{\theta}_0) \xrightarrow{D} N(\mathbf{0}, \mathbf{J}(\boldsymbol{\theta}_0)^{-1}),$$

where $\mathbf{J}(\boldsymbol{\beta}_{10})$ *consists of the first* $s \times s$ *upper-left submatrix of* $\mathbf{J}(\boldsymbol{\beta}_0)$.

Theorem 4.3 establishes the asymptotic properties of $\mathrm{OSE_T}$ under the SCAD penalty. In particular, $\mathrm{OSE_T}$ achieves the same limiting distribution as OSE of $\boldsymbol{\beta}_1$ and $\boldsymbol{\theta}$ in Theorem 3.2 and thus the same efficiency. Furthermore, similar to Theorem 3.2, Theorem 4.3 holds for the class of penalty functions such that $p'_{\lambda_n}(\cdot) = \lambda_n p(\cdot)$ where $p'(\cdot)$ is continuous on $(0, \infty)$, and there is some $\alpha > 0$ such that $p'(\beta) = O(\beta^{-\alpha})$ as $\beta \to 0+$, provided that $N_n^{(1+\alpha)/2}\lambda_n \to \infty$ and $N_n^{1/2}\lambda_n \to 0$.

5. Numerical examples.

5.1. *Simulation study.* We now conduct a simulation study to investigate the finite-sample properties of OSE and $\mathrm{OSE_T}$. The spatial domain of interest is assumed to be a square $[0, l]^2$ of side lengths $l = 5, 10, 15$. The sample sizes are set to be $N = 100, 400, 900$ for $l = 5, 10, 15$, respectively, with a fixed sampling density of 4. For regression, we generate seven covariates that follow standard normal distributions with a cross-covariate correlation of 0.5. The regression coefficients are set to be $\boldsymbol{\beta} = (4, 3, 2, 1, 0, 0, 0)^T$. We standardize the covariates to have mean 0 and variance 1 and standardize $\mathbf{y}$ to have mean 0. Thus, there will be no intercept in the vector of regression coefficients $\boldsymbol{\beta}$. For spatial dependence, the error terms follow an exponential covariance function $\gamma(d) = \sigma^2(1 - c)\exp(-d/r)$, where $\sigma^2 = 9$ is the variance, $c = 0.2$ is the nugget effect and $r = 1$ is the range parameter. For each choice of sample size N, a total of 100 data sets are simulated.

For each simulated data set, we compute OSE and $\mathrm{OSE_T}$ using Algorithm 1. For $\mathrm{OSE_T}$, we consider different threshold values for covariance tapering $\omega = l/2^k$ for $k = 1, 2, \ldots$. We present only the case of $\omega = l/4$ to save space. Our methods are compared against several alternatives. Of particular interest is OSE under a standard linear regression where spatial autocorrelation is unaccounted for in the penalized loglikelihood function. This would be akin to PLS under SCAD in Wang and Zhu [21] and will be referred to as $\mathrm{OSE_{Alt1}}$. In addition, we modify the initialization step of Algorithm 1 by using MLE under the true model which is unknown but assumed to be known. This is an attempt to evaluate the effect of starting values and will be referred to as $\mathrm{OSE_{Alt2}}$. Last, we consider a benchmark case, referred to as $\mathrm{OSE_{Alt3}}$, where the true model is assumed to be known and the MLE of the nonzero regression coefficients and the covariance function parameters

are computed. Our OSE and $\mathrm{OSE}_{\mathrm{T}}$ will be compared against this benchmark to evaluate the oracle properties.

For each choice of sample size N, we first compute the average numbers of correctly (C0) and incorrectly (I0) identified zero-valued regression coefficients from OSE $\widehat{\boldsymbol{\beta}}_{\mathrm{OSE}}$ and $\mathrm{OSE}_{\mathrm{T}}$ $\widehat{\boldsymbol{\beta}}_{\mathrm{OSE}_{\mathrm{T}}}$, as well as those from $\mathrm{OSE}_{\mathrm{Alt1}}$ and $\mathrm{OSE}_{\mathrm{Alt2}}$. The true number of zero-valued regression coefficients is 3 as is assumed in $\mathrm{OSE}_{\mathrm{Alt3}}$. Then, we compute means of the nonzero-valued OSE $\widehat{\boldsymbol{\beta}}_{1,\mathrm{OSE}}$ and $\mathrm{OSE}_{\mathrm{T}}$ $\widehat{\boldsymbol{\beta}}_{1,\mathrm{OSE}_{\mathrm{T}}}$, as well as the corresponding covariance function parameters $\widehat{\boldsymbol{\theta}}_{\mathrm{OSE}}$ and $\widehat{\boldsymbol{\theta}}_{\mathrm{OSE}_{\mathrm{T}}}$. We estimate standard deviations (SDs) of the parameter estimates using the information matrix. The true SD is approximated by the median of the sample SD (SDm) of the 100 parameter estimates. The results are given in Tables 1–3.

In terms of variable selection, C0 tends to the true value 3 and I0 tends to 0, as the sample size N increases, for OSE, $\mathrm{OSE}_{\mathrm{T}}$, $\mathrm{OSE}_{\mathrm{Alt1}}$ and $\mathrm{OSE}_{\mathrm{Alt2}}$. When the sam-

TABLE 1

The average number of correctly identified 0 *coefficients* (*C*0), *average number of incorrectly identified* 0 *coefficients* (*I*0), *mean*, *standard deviation* (*SD*) *and median estimated standard deviation* (*SDm*) *under OSE*, $\mathit{OSE}_{\mathrm{T}}$, $\mathit{OSE}_{\mathrm{Alt1}}$, $\mathit{OSE}_{\mathrm{Alt2}}$ *and* $\mathit{OSE}_{\mathrm{Alt3}}$ *for sample size* $N = 100$

Method	**Truth**	**OSE**	$\mathbf{OSE_T}$	$\mathbf{OSE_{Alt1}}$	$\mathbf{OSE_{Alt2}}$	$\mathbf{OSE_{Alt3}}$
C0	3	2.79	2.84	2.84	2.95	3.00
I0		0.06	0.10	0.32	0.06	0.00
β_1	4.00	4.01	4.03	4.17	4.01	4.01
SD		0.28	0.29	0.39	0.27	0.27
SDm		0.26	0.27	0.36	0.26	0.26
β_2	3.00	3.04	3.03	3.08	3.04	3.03
SD		0.30	0.30	0.41	0.30	0.29
SDm		0.25	0.26	0.36	0.25	0.25
β_3	2.00	1.94	1.97	2.00	1.94	1.93
SD		0.29	0.31	0.50	0.28	0.28
SDm		0.25	0.26	0.36	0.26	0.26
β_4	1.00	1.02	1.03	0.78	1.03	1.02
SD		0.35	0.40	0.55	0.33	0.26
SDm		0.24	0.24	0.26	0.24	0.26
r	1.00	0.79	6.31	–	0.83	0.84
SD		0.54	2.14	–	0.57	0.57
SDm		0.48	17.65	–	0.51	0.51
c	0.20	0.16	0.23	–	0.17	0.17
SD		0.12	0.13	–	0.12	0.12
SDm		0.11	0.19	–	0.11	0.11
σ^2	9.00	7.96	7.14	7.74	8.03	8.03
SD		2.28	1.53	2.06	2.36	2.36
SDm		2.21	4.79	1.16	2.28	2.28

TABLE 2

The average number of correctly identified 0 *coefficients* (*C*0), *average number of incorrectly identified* 0 *coefficients* (*I*0), *mean*, *standard deviation* (*SD*) *and median estimated standard deviation* (*SDm*) *under OSE*, OSE_{T}, OSE_{Alt1}, OSE_{Alt2} *and* OSE_{Alt3} *for sample size* $N = 400$

Method	**Truth**	**OSE**	$\mathbf{OSE_T}$	$\mathbf{OSE_{Alt1}}$	$\mathbf{OSE_{Alt2}}$	$\mathbf{OSE_{Alt3}}$
C0	3	2.97	2.97	2.97	2.98	3.00
I0		0.00	0.00	0.01	0.00	0.00
β_1	4.00	3.98	3.98	3.98	3.99	3.99
SD		0.14	0.14	0.20	0.14	0.14
SDm		0.13	0.13	0.19	0.13	0.13
β_2	3.00	3.02	3.03	3.03	3.02	3.02
SD		0.14	0.14	0.21	0.13	0.13
SDm		0.13	0.13	0.19	0.13	0.13
β_3	2.00	2.01	2.01	2.01	2.01	2.01
SD		0.12	0.12	0.17	0.12	0.12
SDm		0.13	0.13	0.19	0.13	0.13
β_4	1.00	0.99	1.00	0.96	1.00	1.00
SD		0.12	0.12	0.26	0.12	0.12
SDm		0.13	0.13	0.19	0.13	0.13
r	1.00	0.90	2.87	–	0.90	0.90
SD		0.29	4.08	–	0.29	0.29
SDm		0.25	5.24	–	0.25	0.25
c	0.20	0.19	0.29	–	0.19	0.19
SD		0.06	0.07	–	0.06	0.06
SDm		0.05	0.11	–	0.05	0.05
σ^2	9.00	8.70	8.25	8.71	8.70	8.70
SD		1.39	1.00	1.37	1.39	1.39
SDm		1.29	2.95	0.63	1.29	1.29

ple size is relatively small ($N = 100$), $\mathrm{OSE}_{\mathrm{Alt2}}$ has the best performance with the largest C0 and smallest I0, reflecting the effect of starting values in Algorithm 1. But it is not practical, as we do not know what the true model is in actual data analysis. $\mathrm{OSE}_{\mathrm{Alt1}}$ assuming no spatial dependence in the regression model seems to over-shrink the regression coefficients. While $\mathrm{C0} = 2.84$ is close to 3 under $\mathrm{OSE}_{\mathrm{Alt1}}$, $\mathrm{I0} = 0.32$ is also large, compared to our OSE and $\mathrm{OSE}_{\mathrm{T}}$. Between OSE and $\mathrm{OSE}_{\mathrm{T}}$, it appears that C0 is slightly better, but I0 is slightly worse for $\mathrm{OSE}_{\mathrm{T}}$ than OSE.

In terms of estimation of the nonzero regression coefficients, both accuracy and precision improve as the sample size N increases, for all five OSE cases considered here. While the accuracy is similar between $\mathrm{OSE}_{\mathrm{Alt1}}$ and our OSE and $\mathrm{OSE}_{\mathrm{T}}$, a striking feature is the larger SD of $\mathrm{OSE}_{\mathrm{Alt1}}$ when compared with our OSE and $\mathrm{OSE}_{\mathrm{T}}$, for all three sample sizes $N = 100, 400, 900$. This suggests that, by in-

TABLE 3

The average number of correctly identified 0 *coefficients* (*C*0), *average number of incorrectly identified* 0 *coefficients* (*I*0), *mean*, *standard deviation* (*SD*) *and median estimated standard deviation* (*SDm*) *under OSE*, OSE_{T}, OSE_{Alt1}, OSE_{Alt2} *and* OSE_{Alt3} *for sample size* $N = 900$

Method	Truth	OSE	$\mathrm{OSE_T}$	$\mathrm{OSE_{Alt1}}$	$\mathrm{OSE_{Alt2}}$	$\mathrm{OSE_{Alt3}}$
C0	3	3.00	3.00	3.00	3.00	3.00
I0		0.00	0.00	0.00	0.00	0.00
β_1	4.00	4.00	4.01	4.03	4.00	4.00
SD		0.10	0.10	0.13	0.10	0.10
SDm		0.09	0.09	0.13	0.09	0.09
β_2	3.00	3.01	3.01	2.99	3.01	3.01
SD		0.08	0.08	0.12	0.08	0.08
SDm		0.09	0.09	0.13	0.09	0.09
β_3	2.00	1.98	1.99	1.98	1.98	1.98
SD		0.08	0.08	0.11	0.08	0.08
SDm		0.09	0.09	0.13	0.09	0.09
β_4	1.00	1.00	1.00	1.01	1.00	1.00
SD		0.09	0.09	0.13	0.09	0.09
SDm		0.09	0.09	0.13	0.09	0.09
r	1.00	0.94	1.44	–	0.94	0.94
SD		0.17	0.50	–	0.17	0.17
SDm		0.17	0.40	–	0.17	0.17
c	0.20	0.19	0.25	–	0.19	0.19
SD		0.04	0.04	–	0.04	0.04
SDm		0.04	0.04	–	0.04	0.04
σ^2	9.00	8.80	8.50	8.80	8.80	8.80
SD		0.90	0.74	0.87	0.90	0.90
SDm		0.85	1.15	0.42	0.85	0.85

cluding spatial dependence directly in the penalized likelihood function, we gain statistical efficiency in parameter estimation. For the small sample size ($N = 100$), SD based on the information matrix without accounting for spatial dependence appears to underestimate the true variation estimated by SDm. Furthermore, the SD's of OSE and $\mathrm{OSE_T}$ tend to those in the benchmark case $\mathrm{OSE_{Alt3}}$ as the sample size increases, confirming the oracle properties established in Sections 3 and 4. For 100 simulations, it takes about 1 second, 30 seconds and 4 minutes per simulation for sample sizes $N = 100$, 400, 900, respectively.

Based on these simulation results, it may be tempting to consider using $\mathrm{OSE_{Alt1}}$ to select variables and then $\mathrm{OSE_{Alt3}}$ for parameter estimation when the sample size is reasonably large, as a means of saving computational time. We contend that this is not necessary, as our OSE or $\mathrm{OSE_T}$ enables variable selection and parameter estimation simultaneously, at the similar computational cost. Moreover, in prac-

tice, it is not always clear how large a sample size at hand really is, as an effective sample size is influenced by factors such as the strength of spatial dependence in the error process.

5.2. *Data examples*. The first data example consists of January precipitation (inches per 24-hour period) on the log scale from 259 weather stations in the state of Colorado [15]. Candidate covariates are elevation, slope, aspect and seven spectral bands from a MODIS satellite imagery (B1M through B7M). It is of interest to investigate the relationship between precipitation and these covariates.

We first fit a spatial linear model with an exponential covariance function via maximum likelihood. The parameter estimates and their standard errors in Table 4 suggest that the regression coefficients for elevation, B1M, B4M, B6M and B7M are possibly significant. Among the covariance function parameters, of most interest is the range parameter, which is significantly different from zero. This indicates that there is spatial autocorrelation among the errors in the linear regression. Our OSE method selects elevation and B4M, and shrinks all the other regression coefficients to zero. The covariance function parameter estimates are close to the MLE. For comparison, we fit a standard linear regression with i.i.d. errors and the corresponding OSE_{Alt1} selects slope and aspect in addition to elevation and B4M. However, the regression coefficients for slope and aspect do not appear to be significant.

TABLE 4

Regression coefficient estimates and standard deviations (SD) using maximum likelihood (MLE) and one-step sparse estimation (OSE) under a spatial linear model with an exponential covariance function for the Gaussian error process, as well as OSE under a standard linear model with i.i.d. errors (OSE_{Alt1})

Terms	**MLE**	**SD**	**OSE**	**SD**	$\mathbf{OSE_{Alt1}}$	**SD**
			Regression coefficients			
Elevation	0.305	0.055	0.228	0.054	0.195	0.044
Slope	0.016	0.026	–	–	0.035	0.040
Aspect	−0.004	0.022	–	–	0.032	0.034
B1M	0.214	0.157	–	–	–	–
B2M	0.058	0.064	–	–	–	–
B3M	0.017	0.109	–	–	–	–
B4M	−0.404	0.183	−0.089	0.034	−0.264	0.045
B5M	0.043	0.089	–	–	–	–
B6M	−0.162	0.116	–	–	–	–
B7M	0.172	0.098	–	–	–	–
			Covariance function parameters			
Range	0.967	0.368	1.043	0.417	–	–
Nugget	0.183	0.061	0.196	0.064	–	–
σ^2	0.287	0.067	0.304	0.074	0.289	0.026

In addition, we apply our method to the whiptail lizard data as described in Section 1. There are 148 sites, and the response variable is the abundance of lizards at each site. There are 26 covariates regarding location, vegetation, flora, soil and ants. Hoeting et al. [10] considered only 6 covariates after a separate prescreening procedure, and selected 2 covariates in their final model. In this paper, we consider all 26 covariates simultaneously, and interestingly reach the same final model. For details of the results, see Section 6.2 in Chu et al. [1].

APPENDIX: TECHNICAL DETAILS

For ease of notation, we suppress n in ${}^n t_{kk'}$, ${}^n a_{kk'}$, ${}^n\boldsymbol{\Gamma}$, $\mathbf{I}_n$, $\mathbf{A}_n$, ${}^n\widehat{\boldsymbol{\eta}}$, ${}^n\widehat{\boldsymbol{\beta}}$ and ${}^n\widehat{\boldsymbol{\theta}}$. The detailed proofs of all lemmas and theorems are given in Chu et al. [1].

A.1. Asymptotic properties of PMLE and OSE.

LEMMA 1. *Under* (A.1)–(A.7), *for any given* $\boldsymbol{\eta} \in \mathbb{R}^p \times \Omega$, *we have, as* $n \to \infty$,

$$N_n^{-1/2}\ell'(\boldsymbol{\eta}) \xrightarrow{D} N(\mathbf{0}, \mathbf{J}(\boldsymbol{\eta})), \qquad N_n^{-1}\ell''(\boldsymbol{\eta}) \xrightarrow{P} -\mathbf{J}(\boldsymbol{\eta}),$$

where $\mathbf{J}(\boldsymbol{\eta}) = \operatorname{diag}\{\mathbf{J}(\boldsymbol{\beta}), \mathbf{J}(\boldsymbol{\theta})\}$.

REMARK. Lemma 1 establishes the asymptotic behavior of the first-order and the second-order derivatives of the log-likelihood function $\ell(\boldsymbol{\eta})$, scaled by $N_n^{-1/2}$ and N_n^{-1}, respectively. In addition, by Theorem 2 of Mardia and Marshall [14], $\widehat{\boldsymbol{\eta}}_{\mathrm{MLE}}$ is consistent and asymptotically normal with $\|\widehat{\boldsymbol{\eta}}_{\mathrm{MLE}} - \boldsymbol{\eta}_0\| = O_p(N_n^{-1/2})$ and $N_n^{1/2}(\widehat{\boldsymbol{\eta}}_{\mathrm{MLE}} - \boldsymbol{\eta}_0) \xrightarrow{D} N(\mathbf{0}, \mathbf{J}(\boldsymbol{\eta}_0)^{-1})$. Moreover, for a random vector $\boldsymbol{\eta}^*$, such that $\|\mathbf{I}(\boldsymbol{\eta})^{1/2}(\boldsymbol{\eta}^* - \boldsymbol{\eta})\| = O_p(1)$, by Theorem 2 of Mardia and Marshall [14], we have $N_n^{-1}\ell''(\boldsymbol{\eta}^*) \xrightarrow{P} -\mathbf{J}(\boldsymbol{\eta})$. These results will be used repeatedly in the proof of Theorems 3.1 and 3.2.

PROOF OF THEOREM 3.1. The proof follows from Lemma 1 and arguments extended from Theorems 1 and 2 in Fan and Li [7]. See details in Chu et al. [1]. □

PROOF OF THEOREM 3.2. The proof follows from Lemma 1 and arguments extended from Theorem 5 in Zou and Li [24]. See details in Chu et al. [1]. □

A.2. Asymptotic properties of $\mathrm{PMLE}_{\mathrm{T}}$ and $\mathrm{OSE}_{\mathrm{T}}$. Let $|A|$ denote the cardinality of a discrete set A. Let $\mu_{1,\mathrm{T}} \le \cdots \le \mu_{N_n,\mathrm{T}}$ denote the eigenvalues of tapered covariance matrix $\boldsymbol{\Gamma}_{\mathrm{T}}$. Let $\mu_{l,\mathrm{T}}^k$ denote the eigenvalues of $\boldsymbol{\Gamma}_{k,\mathrm{T}}$ such that $|\mu_{1,\mathrm{T}}^k| \le \cdots \le |\mu_{N_n,\mathrm{T}}^k|$ and let $\mu_{l,\mathrm{T}}^{kk'}$ denote the eigenvalues of $\boldsymbol{\Gamma}_{kk',\mathrm{T}}$ such that $|\mu_{1,\mathrm{T}}^{kk'}| \le \cdots \le |\mu_{N_n,\mathrm{T}}^{kk'}|$. For a matrix $\mathbf{A}$, we let $\mu_{\min}(\mathbf{A})$ denote the minimum eigenvalue of $\mathbf{A}$. Also, recall that $t_{kk',\mathrm{T}} = \operatorname{tr}(\boldsymbol{\Gamma}_{\mathrm{T}}^{-1}\boldsymbol{\Gamma}_{k,\mathrm{T}}\boldsymbol{\Gamma}_{\mathrm{T}}^{-1}\boldsymbol{\Gamma}_{k',\mathrm{T}})$.

LEMMA 2. *Under* (A.12)–(A.13), *we have*:

(i) $\|\mathbf{\Gamma} - \mathbf{\Gamma}_{\mathrm{T}}\|_\infty = O(N_n^{-1/2})$; (ii) $\|\mathbf{\Gamma}_k - \mathbf{\Gamma}_{k,\mathrm{T}}\|_\infty = O(N_n^{-1/2})$;

(iii) $\|\mathbf{\Gamma}_{kk'} - \mathbf{\Gamma}_{kk',\mathrm{T}}\|_\infty = O(N_n^{-1/2})$.

REMARK. Lemma 2 establishes that the order of the difference between the covariance matrix $\mathbf{\Gamma}$ and the tapered covariance matrix $\mathbf{\Gamma}_{\mathrm{T}}$ is $N_n^{-1/2}$, as well as that of the first-order and the second-order derivatives of the covariance matrices. These results are used when establishing Lemma 3.

LEMMA 3. *Under* (A.1)–(A.4), (A.6) *and* (A.12), (A.13), *we have*:

(C.1) $\lim_{n\to\infty} \mu_{N_n,\mathrm{T}} = C < \infty, \lim_{n\to\infty} |\mu^k_{N_n,\mathrm{T}}| = C_k < \infty, \lim_{n\to\infty} |\mu^{kk'}_{N_n,\mathrm{T}}| = C_{kk'} < \infty$ *for any* $k, k' = 1, \ldots, q$.

(C.2) *For* $k = 1, \ldots, q$, $\|\mathbf{\Gamma}_{k,\mathrm{T}}\|_F^{-2} = O(N_n^{-1/2-\delta})$, *for some* $\delta > 0$.

(C.3) $\|\mathbf{\Gamma}_{\mathrm{T}}^{-1}\|_s < C_0 < \infty$.

(C.4) *For any* $k, k' = 1, \ldots, q$, $a_{kk',\mathrm{T}} = \lim\{t_{kk',\mathrm{T}}(t_{kk,\mathrm{T}} t_{k'k',\mathrm{T}})^{-1/2}\}$ *exists and is equal to* $a_{kk'} = \lim\{t_{kk'}(t_{kk} t_{k'k'})^{-1/2}\}$. *That is,* $\mathbf{A}_{\mathrm{T}} = (a_{kk',\mathrm{T}})^q_{k,k'=1} = \mathbf{A} = (a_{kk'})^q_{k,k'=1}$ *and is nonsingular.*

REMARK. Conditions (C.1)–(C.4) are the covariance tapering counterparts of (A.2), (A.3)(i), (A.4)(i) and (A.6). Together with (A.5), they yield Proposition 4.1. In fact, Lemmas 2 and 3 hold for other tapering functions such as truncated polynomial functions of d/ω with constant term equal to 1 when $d < \omega$, and 0 otherwise [22]. Furthermore, (A.12) can be weakened to $0 < \inf_n\{\omega_n N_n^{-1/2+\tau}\} \le \sup_n\{\omega_n N_n^{-1/2+\tau}\} < \infty$, with $\tau < \min\{1/2, \delta\}$.

LEMMA 4. *Under* (A.1)–(A.7) *and* (A.12)–(A.13), *for any given* $\boldsymbol{\eta} \in \mathbb{R}^p \times \Omega$, *we have*

$$N_n^{-1/2} \ell'_{\mathrm{T}}(\boldsymbol{\eta}) \xrightarrow{D} N(\mathbf{0}, \mathbf{J}(\boldsymbol{\eta})) \quad \text{and} \quad N_n^{-1} \ell''_{\mathrm{T}}(\boldsymbol{\eta}) \xrightarrow{P} -\mathbf{J}(\boldsymbol{\eta}),$$

where recall that $\mathbf{J}(\boldsymbol{\eta}) = \mathrm{diag}\{\mathbf{J}(\boldsymbol{\beta}), \mathbf{J}(\boldsymbol{\theta})\}$.

REMARK. Lemma 4 establishes the asymptotic behavior of the first-order and the second-order derivatives of the covariance-tapered log-likelihood function $\ell_{\mathrm{T}}(\boldsymbol{\eta})$. The rates of convergence and the limiting distributions are the same as those for the log-likelihood function. As in Lemma 1, it follows that $\mathrm{MLE}_{\mathrm{T}}$ $\widehat{\boldsymbol{\eta}}_{\mathrm{MLE}_{\mathrm{T}}}$ is consistent and asymptotically normal, as is given in Proposition 4.1. These results will be used to establish Theorems 4.2 and 4.3 and play the same role as Lemma 1 when showing Theorems 3.1 and 3.2.

PROOF OF PROPOSITION 4.1. From Lemma 3, (C.1)–(C.4) are satisfied. Together with (A.5), the regularity conditions of Theorem 2 of Mardia and Marshall [14] hold. Thus, the result in Proposition 4.1 follows. □

PROOF OF THEOREM 4.2. The proof of Theorem 4.2 is similar to that of Theorem 3.1. The main differences are that the parameter estimates $\widehat{\boldsymbol{\eta}}_{\mathrm{PMLE}}$, log-likelihood function $\ell(\boldsymbol{\eta})$ and penalized log-likelihood $Q(\boldsymbol{\eta})$ are replaced with their covariance-tapered counterparts $\widehat{\boldsymbol{\eta}}_{\mathrm{PMLE_T}}$, $\ell_{\mathrm{T}}(\boldsymbol{\eta})$ and $Q_{\mathrm{T}}(\boldsymbol{\eta})$, respectively. Furthermore, we replace the results from Lemma 1 with those from Lemma 4, which holds due to Lemmas 2 and 3 under the additional assumptions (A.12) and (A.13). □

PROOF OF THEOREM 4.3. The proof of Theorem 4.3 is similar to that of Theorem 3.2, but we replace the parameter estimates $\widehat{\boldsymbol{\eta}}_{\mathrm{OSE}}$, log-likelihood function $\ell(\boldsymbol{\eta})$ and $Q^*(\boldsymbol{\beta})$ with their covariance-tapered counterparts $\widehat{\boldsymbol{\eta}}_{\mathrm{OSE_T}}$, $\ell_{\mathrm{T}}(\boldsymbol{\eta})$ and $Q^*_{\mathrm{T}}(\boldsymbol{\beta})$, respectively. As before, we replace the results from Lemma 1 with those from Lemma 4, where the additional conditions (A.12) and (A.13) are assumed and Lemmas 2 and 3 are applied. □

Acknowledgments. We are grateful to the Editor, Associate Editor and three anonymous referees for their helpful and constructive comments. We thank Drs. Jennifer Hoeting and Jay Ver Hoef for providing the lizard data.

REFERENCES

[1] CHU, T., ZHU, J. and WANG, H. (2011). Penalized maximum likelihood estiamtion and variable selection in geostatistics. Technical report, Dept. Statistics, Colorado State Univ., Fort Collins, CO.

[2] CRESSIE, N. A. C. (1993). *Statistics for Spatial Data*, revised ed. Wiley, New York. MR1239641

[3] DRAPER, N. R. and SMITH, H. (1998). *Applied Regression Analysis*, 3rd ed. Wiley, New York. MR1614335

[4] DU, J., ZHANG, H. and MANDREKAR, V. S. (2009). Fixed-domain asymptotic properties of tapered maximum likelihood estimators. *Ann. Statist.* **37** 3330–3361. MR2549562

[5] EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. MR2060166

[6] FAN, J. (1997). Comments on "Wavelets in statistics: A review," by A. Antoniadis. *Journal of the Italian Statistical Association* **6** 131–138.

[7] FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

[8] FAN, J. and PENG, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32** 928–961. MR2065194

[9] FURRER, R., GENTON, M. G. and NYCHKA, D. (2006). Covariance tapering for interpolation of large spatial datasets. *J. Comput. Graph. Statist.* **15** 502–523. MR2291261

[10] HOETING, J. A., DAVIS, R. A., MERTON, A. A. and THOMPSON, S. E. (2006). Model selection for geostatistical models. *Ecol. Appl.* **16** 87–98.

[11] HUANG, H.-C. and CHEN, C.-S. (2007). Optimal geostatistical model selection. *J. Amer. Statist. Assoc.* **102** 1009–1024. MR2411661

[12] KAUFMAN, C. G., SCHERVISH, M. J. and NYCHKA, D. W. (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. *J. Amer. Statist. Assoc.* **103** 1545–1555. MR2504203

[13] LINHART, H. and ZUCCHINI, W. (1986). *Model Selection*. Wiley, New York. MR0866144

[14] MARDIA, K. V. and MARSHALL, R. J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* **71** 135–146. MR0738334

[15] REICH, R. and DAVIS, R. (2008). *Lecture Notes of Quantitative Spatial Analysis*. Colorado State University, Fort Collins, CO.

[16] SCHABENBERGER, O. and GOTWAY, C. A. (2005). *Statistical Methods for Spatial Data Analysis*. Chapman and Hall/CRC, Boca Raton, FL. MR2134116

[17] STEIN, M. L. (1999). *Interpolation of Spatial Data*: *Some Theory for Kriging*. Springer, New York. MR1697409

[18] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

[19] WANG, H., LI, G. and TSAI, C.-L. (2007). Regression coefficient and autoregressive order shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69** 63–78. MR2301500

[20] WANG, H., LI, R. and TSAI, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94** 553–568. MR2410008

[21] WANG, H. and ZHU, J. (2009). Variable selection in spatial regression via penalized least squares. *Canad. J. Statist.* **37** 607–624. MR2588952

[22] WENDLAND, H. (1995). Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Adv. Comput. Math.* **4** 389–396. MR1366510

[23] ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

[24] ZOU, H. and LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36** 1509–1533. MR2435443

T. CHU
H. WANG
DEPARTMENT OF STATISTICS
COLORADO STATE UNIVERSITY
FORT COLLINS, COLORADO 80523
USA
E-MAIL: tingjin.chu@colostate.edu
wanghn@stat.colostate.edu

J. ZHU
DEPARTMENT OF STATISTICS
AND DEPARTMENT OF ENTOMOLOGY
UNIVERSITY OF WISCONSIN
MADISON, WISCONSIN 53706
USA
E-MAIL: jzhu@stat.wisc.edu