Spatial Statistics 25 (2018) 86–104

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta

# A penalized quasi-maximum likelihood method for variable selection in the spatial autoregressive model

Xuan Liu [a,b], Jianbao Chen [a,*], Suli Cheng [a]

[a] College of Mathematics and Informatics, Fujian Normal University, Fuzhou, 350117, PR China
[b] Department of Basic Teaching and Research, Yango University, Fuzhou, 350015, PR China

ARTICLE INFO

*Article history:*
Received 4 December 2017
Accepted 7 May 2018

*Keywords:*
Spatial autoregressive model
Variable selection
Penalized quasi-maximum likelihood
SCAD
Oracle property

ABSTRACT

This paper investigates variable selection in the spatial autoregressive model with independent and identical distributed errors. A penalized quasi-maximum likelihood method is developed for simultaneous model selection and parameter estimation. Under some regular conditions, theoretical properties of the proposed estimators, including consistency and the oracle property, are established. In addition, a computationally feasible algorithm is designed to realize variable selection procedure. Simulation studies are conducted to examine the finite sample performance of the proposed method and a real example about Boston housing data is presented for illustration purpose.

## 1. Introduction

Spatial regression models are important tools in dealing with spatial dependent data which is widely available in many fields such as spatial econometrics and regional science. Among them, the spatial autoregressive (SAR) model proposed by Cliff and Ord (1973) has received much attention; see, e.g., the research by Anselin and Bera (1998), and the books by Anselin (1988) and Cressie (1993). These studies mainly focused on estimation of unknown parameters in the models. When the number of covariates in the SAR model is large, how to select significant variables to enhance predictability and reduce calculation amount becomes a very important problem. However, variable selection in spatial regression models is more complex and difficult due to spatial dependence compared with classical

[*] Corresponding author.
*E-mail address:* jbjy2006@126.com (J. Chen).

https://doi.org/10.1016/j.spasta.2018.05.001
2211-6753/© 2018 Elsevier B.V. All rights reserved.

linear regression models. As a result, some important penalized methods for variable selection, such as the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996), the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001), cannot be used directly in the SAR model and the related properties have not been studied in this respect. Therefore, we propose a penalized method of variable selection for the SAR model and investigate its properties thoroughly. Furthermore, an efficient algorithm is designed to complete this procedure.

Variable selection for the classical linear models has been discussed by many researchers. There are two main kinds of methods in traditional analysis. One is hypothesis testing such as F-tests in a stepwise selection procedure (Draper and Smith, 1998); the other is to select models based on information criterion such as Akaike information criterion (AIC), Bayesian information criterion (BIC), risk inflation criterion (RIC) and etc. Although they are practically useful, the common drawback is that one need compare all possible submodels. This is a combinational problem with NP-complexity (Huo and Ni, 2007). In recent years, penalized methods of variable selection have attracted a great deal of research, including LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), elastic-net (ENet) (Zou and Hastie, 2005), adaptive LASSO (ALASSO) (Zou, 2006), minimax concave penalty (MCP) (Zhang, 2010) and so on. These methods can estimate regression coefficients and select the best model simultaneously. Under mild regular conditions, Fan and Li (2001) firstly established the asymptotic properties of penalized likelihood estimators and in particular, the oracle property in the sense that the efficiency of an estimator behaves the same as the oracle ordinary least squares estimator under the true model (Donoho and Johnstone, 1994). More detailed information can be found in Chen et al. (2014). The introduction of other different methods, such as Bayesian model averaging and frequentist model averaging, were presented by Steel (2017).

Recent studies of variable selection in spatial data have been receiving increasing attention. Based on AIC, Hoeting et al. (2006) established variable selection procedure for geostatistical data, which can be easy to get into time-consuming calculations when a large number of covariates are included in the original model. Thus, popular penalized methods were introduced in this case. Generally, it is challenging to extend the penalized methods to data that are dependent either over time or across space, as variable selection involves not only regression coefficients but also auto-correlation coefficients (Zhu et al., 2010). There are some developments in spatial dependent data. Huang et al. (2006) provided a modification of the LASSO procedure that simultaneously performs variable selection, spatial neighbourhood selection and parameter estimation in spatial lattice data. Wang and Zhu (2009) established a penalized least squares under various penalty functions for a spatial linear model where the error process is assumed to be a second-order stationary random field. Zhu et al. (2010) studied an adaptive LASSO for selection both covariates and neighbourhood structures in spatial linear models with Gaussian process errors. Chu et al. (2011) developed a penalized maximum likelihood estimation to select covariates as well as parameter estimation simultaneously in spatial linear models with Gaussian process errors. Those studies established the theoretical properties of their methods, such as oracle property.

In this paper, we consider variable selection for the most popular spatial autoregressive (SAR) model in regular lattice data. Note that the SAR model contains the spatial lagged dependent variable while the spatial models listed above do not introduce it. Starting with the seminal work on Bayesian model selection for the SAR model by LeSage and Parent (2007), there is a great bulk of literature both on variable as well as model selection for SAR models. LeSage and Parent (2007) developed the Markov Chain Monte Carlo model composition methodology ($MC^3$) and Bayesian model averaging (BMA) technique for the SAR and spatial error models, and focused exclusively on model specification issues regarding the choice of explanatory variables as in conventional linear models. Subsequently, some extensions of the work of LeSage and Parent (2007) include LeSage and Fischer (2008) and Cuaresma et al. (2014, 2018). Since BMA techniques for the SAR models rely on the calculation of marginal likelihoods, there is a severe computational burden when a large number of covariates are potential candidates of the specification. To reduce the time consuming, Piribauer and Fischer (2015) proposed the use of posterior model weights based on the Bayesian information criterion and maximum likelihood estimates of the matrix exponential specification (see LeSage and Parent, 2007) of global spatial spill over effects. Piribauer (2016) used stochastic search variable selection (SSVS) priors to deal with the problem of variable selection in the SAR models, which can avoid the complex calculation