Computational Statistics and Data Analysis 87 (2015) 116–135

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Adjusted quasi-maximum likelihood estimator for mixed regressive, spatial autoregressive model and its small sample bias

Dalei Yu *, Peng Bai, Chang Ding

Statistics and Mathematics College, Yunnan University of Finance and Economics, Kunming 650221, China
Yunnan Tongchuang Scientific Computing & Data Mining Center, Kunming 650221, China

ARTICLE INFO

Article history:
Received 22 February 2014
Received in revised form 3 December 2014
Accepted 3 February 2015
Available online 12 February 2015

Keywords:
Adjusted quasi-maximum likelihood estimator
Degrees of freedom loss
Mixed regressive
Spatial autoregressive
Small sample bias

ABSTRACT

Under flexible distributional assumptions, the adjusted quasi-maximum likelihood (ADQML) estimator for mixed regressive, spatial autoregressive model is studied in this paper. The proposed estimation method accommodates the extra uncertainty introduced by the unknown regression coefficients. Moreover, the explicit expressions of theoretical/feasible second-order-bias of the ADQML estimator are derived and the difference between them is investigated. The feasible second-order-bias corrected ADQML estimator is then designed accordingly for small sample setting. Extensive simulation studies are conducted under both normal and non-normal situations, showing that the quasi-maximum likelihood (QML) estimator suffers from large bias when the sample size is relatively small in comparison to the number of regression coefficients and such bias can be effectively eliminated by the proposed ADQML estimation method. The use of the method is then demonstrated in the analysis of the Neighborhood Crimes Data.

1. Introduction

Mixed regressive, spatial autoregressive (MRSAR) model is widely used in geostatistics, spatial econometrics, regional science and urban economics (Lee, 2004; Anselin, 2010). Earlier studies about parameter estimation and computational problem can be found in Ord (1975) and Anselin (1988). Moreover, the large sample properties of the quasi-maximum likelihood (QML) estimator were studied in Lee (2004) where the estimator is derived from a hypothetical normal likelihood but the errors (disturbances) in the model are not truly normally distributed. Other estimation and hypothesis testing problems were studied in Kelejian and Prucha (1999, 2001) and Lee (2007), among others. The model was further developed by Lee and Yu (2010), Baltagi et al. (2012) and Millo (2014), where the spatial panel model and random effects model were considered.

Following Anselin (1988) and Lee (2004), the MRSAR model is defined as

$$Y_n = X_n\beta + \rho W_n Y_n + V_n, \tag{1}$$

where n is the total number of spatial units, $Y_n = (y_1, \dots, y_n)^{\mathrm{T}}$ is the $n \times 1$ observable responses vector, X_n is the $n \times k$ matrix of regressors, W_n is a specified constant spatial weight matrix with elements $\{w_{ij}, i, j = 1, \dots, n\}$ representing the "degrees of possible interaction" of location j on location i and $w_{ii} = 0$ (Ord, 1975), $V_n = (v_1, \dots, v_n)^{\mathrm{T}}$ is the n-dimensional vector of random errors (disturbances) with zero mean. The autoregressive coefficient ρ is the scaling parameter which captures the "interaction" between y_i and the weighted average of the neighbors y_j (i.e. spatial lag).

* Corresponding author at: Statistics and Mathematics College, Yunnan University of Finance and Economics, Kunming 650221, China.
E-mail addresses: yudalei@126.com (D. Yu), baipeng68@126.com (P. Bai), dingchang81@gmail.com (C. Ding).

http://dx.doi.org/10.1016/j.csda.2015.02.003
0167-9473/

Let $S_n(\rho) = (I_n - \rho W_n)$, if $V_n \sim N(0_{n\times 1}, \sigma^2 I_n)$, then the log-likelihood function of $\xi = (\beta^{\mathrm{T}}, \rho, \sigma^2)^{\mathrm{T}}$ given Y_n is

$$L_n(\xi) = -\frac{n \log(2\pi\sigma^2)}{2} + \log\det S_n(\rho) - \frac{\|S_n(\rho)Y_n - X_n\beta\|_2^2}{2\sigma^2}, \tag{2}$$

where $\|a\|_2^2 = a^{\mathrm{T}}a$ for vector a. When V_n is not truly normally distributed, $L_n(\xi)$ is still adopted as the criterion function and is termed as the log quasi-likelihood function (Lee, 2004). The resulting estimator by maximizing this log quasi-likelihood function is referred to as the quasi-maximum likelihood (QML) estimator. Lee (2004) showed that under certain regularity conditions the QML estimator has important asymptotic properties (for example the $n^{1/2}$-consistency, asymptotic normality and efficiency when model is correctly specified). Essentially, the autoregressive coefficient ρ plays a central role in the spatial data analysis. However, the regression coefficients vector β and ρ, σ^2 are "tied up" together and therefore the uncertainty caused by estimating β will be introduced to the estimation procedure of ρ and σ^2 and leads to specific bias. Such bias becomes apparent when the number of regression coefficients k is large in comparison with n and some literatures term this phenomenon as "degrees of freedom loss" (Harville, 1977; Jiang, 2007). Please refer Corollary 2 and Simulations 1 and 2 for more details.

The deficiency caused by "degrees of freedom loss" can be effectively relieved by adjusting the concentrated or profile likelihood/score function (Harville, 1977; Durban and Currie, 2000; Jiang, 2007). The resulting estimator sometimes is termed as restricted maximum likelihood (REML) estimator and such estimator has been studied in different literatures (see Jiang, 2007 and the references therein for details). However, in the existing studies for REML estimator, the major research focus is on developing the corresponding large sample properties and these studies are usually based on specific distributional assumptions (Cressie and Lahiri, 1993; Jiang, 2007). In fact, in some situations it is necessary to consider the small sample properties of the estimators (for example in Section 4.2, the sample size $n = 49$).

Focusing on the small sample properties of ML estimator for pure spatial autoregressive model, Bao and Ullah (2007) studied its second-order-bias and mean square error, where no exogenous regressors are involved and V_n is assumed to be normally distributed. Moreover, Bao (2013) studied the explicit form of finite sample bias (up to second-order) of QML estimators in MRSAR model and corrected the second-order-bias directly based on the original full score function. However, Bao's formulation essentially is based on QML-type score function and thus the "degrees of freedom loss" effect will impose difficulties when different components of QML estimators have different convergence rates (see the discussion in Section 3 of Bao, 2013 for details). Moreover, Bao's bias expression involves high dimensional matrix manipulation and thus it is difficult to study its analytical properties (Bao, 2013; Yang, 2014). Recently, Yang (2014) proposed a general method to evaluate the second/third-order-bias via semi-parametric bootstrap. The method is flexible and does not require the closed form of bias. However, in some situations the explicit form of the bias is also of research interest. The reason is that such expression allows us to calculate the bias over the parameter space, which can provide a valuable source of information (Bao, 2013). Moreover, based on the explicit bias expression, one can design the corresponding bias corrected estimator in a computationally attractive way, which can be used conveniently in the situation where many models are involved (e.g. in model comparison and model averaging). As such, it is necessary to adjust the existing QML-type formulation appropriately to address these problems and investigate the corresponding large/small sample properties analytically, especially when the complete distribution of V_n is unknown.

Under flexible distributional assumptions (only basic moment assumptions on V_n are imposed) we will handle these aforementioned concerns in this paper. Current study has two contributions. First, under a straightforward framework, we provide the expression of an adjusted QML (ADQML) estimator to overcome the "degrees of freedom loss" caused by estimating β. Second, focusing on the small sample setting, the explicit form of second-order-bias of ADQML estimator is derived and the corresponding properties are assessed analytically. The rest of the paper is arranged as follows: Section 2 provides the formulation of ADQML estimator and the large sample properties of the ADQML estimator are investigated in the light of the results given by Lee (2004). The explicit expression of small sample bias (up to second-order) of the ADQML estimator, the corresponding bias corrected estimator and its analytical properties are studied in Section 3. Simulation studies are conducted in Section 4 to assess the performance of the proposed methods in finite sample settings and the use of the methods are demonstrated in the analysis of the Neighborhood Crimes Data. The last section provides the concluding discussions and some possible further research directions. Proofs are provided in the Appendix.

2. Formulation of ADQML estimator

It is clear that QML estimators are obtained by solving

$$\frac{1}{n}\frac{\partial L_n(\xi)}{\partial \xi} = \frac{1}{n}\begin{pmatrix} \frac{1}{\sigma^2}X_n^{\mathrm{T}}\left\{S_n(\rho)Y_n - X_n\beta\right\} \\ -\mathrm{tr}\left\{G_n(\rho)\right\} + \frac{\left\{S_n(\rho)Y_n - X_n\beta\right\}^{\mathrm{T}}W_nY_n}{\sigma^2} \\ -\frac{n}{2\sigma^2} + \frac{\|S_n(\rho)Y_n - X_n\beta\|_2^2}{2\left(\sigma^4\right)} \end{pmatrix} = 0_{(k+2)\times 1}. \tag{3}$$