Bootstrap Test Statistics for Spatial Econometric Models

Kuan-Pin Lin*
Portland State University

Zhi-He Long† Wu Mei
South China University of Technology

Current Version: April 22, 2007

Abstract

We introduce and apply bootstrap method for testing spatial correlation in a linear regression model. Given the consideration of a fixed spatial structure and heteroscedasticity of unknown form in the data, spatial bootstrap is a hybrid of recursive wild bootstrap method. Based on the Moran’s index I, two versions of LM-Error and LM-Lag statistics, we demonstrate that the spatial bootstrap procedure can be used for model identification (pre-test) and diagnostic checking (post-test) of a spatial econometric model. With two empirical examples, the bootstrap method is proven to be an effective alternative to the theoretical asymptotic approach for hypothesis testing when the distribution assumption is violated.


JEL Classification: C12, C15, C21, R11

* Kuan-Pin Lin, Department of Economics, Portland State University, Portland, Oregon, USA, 97207. Tel: 001+503-725-3931. E-Mail: link@pdx.edu
† Zhi-He Long, School of Business Administration, South China University of Technology, Guangzhou, China, 510640. Tel: 86+20-8711-4048. E-Mail: l-zh720@163.com
1. Introduction

The rapid development of spatial econometrics began in about two decades ago (Anselin [1988], See also Anselin and Bera [1997] for a survey). It studies the effects of spatial structure and interaction in cross section and panel data. More recent applications have extended to combine the time series dynamics in a spatial framework (see Baltagi, [2003], Elhorst [2006], Kapoor, Keleijian, and Prucha [2006]). Fueled by the breakthrough of GIS hardware and spatial statistical software, we foresee continuing development of spatial econometric methods in years to come.

On the theoretical front, the asymptotic theory of the test statistics for spatial model specification has been the investigation focus since Cliff and Ord (1972, 1973, see also Sen [1976], King [1981], Anselin [1988], Anselin and Rey [1991], Anselin, Bera, Florax, and Yoon [1996], Anselin and Florax [1995]). Only until recently, the general asymptotic treatment of Moran-based test statistics has become available (Anselin and Keleijian [1997], Keleijian and Prucha [2001], Pinkse [1999, 2004]). The theory confirms that limited distributions exist for Moran-based statistics for testing spatial dependence under a set of regularity conditions for the i.i.d. normal model error structure. For empirical applications, however, it is rare that the theoretical requirements of the model will be satisfied. Theory may serve the ideal condition for model construction, but its usefulness in deriving the realistic interpretation is questionable, in particular with a small sample. In this paper, we introduce a bootstrap method for diagnostic testing spatial correlation in data and model specification. We believe that empirical evidence often in small sample could be bootstrapped to emulate the theory in large sample.

In this paper, we apply bootstrap method for hypothesis testing spatial in a linear regression model with spatial correlation and general heteroscedasticity. The particular method we advocated is a hybrid version of residual-based recursive wild bootstrap (see Wu [1986], Liu [1988], Davidson and Flachaire [2001], and Goncalves and Kilian [2004] for a time series application). The bootstrap method for specification search begins from a pre-test of a linear regression model under null hypothesis of spatial independence to diagnostic checking or post-testing of the final model. The pre-test is typically performed on a basic linear regression model in searching for a proper spatial model specification.
Given a certain spatial structure identified, the modified model is a non-linear regression model that subject to diagnostic testing. The model is estimated with quasi-maximum log-likelihood method with the concern that the model is potentially misspecified. We bootstrap five Moran-based test statistics popular in the spatial econometric literature within a consistent framework for both linear and nonlinear regression models. The test statistics are not exact but they are asymptotic pivotal\(^1\), therefore asymptotic valid for the application of bootstrap methods (see MacKinnon [2002] for a review on the asymptotic theory of bootstrap). Theoretical investigation of the size and power of spatial bootstrap method is beyond the scope of this paper, and will be an important topic for future research.

In the next Section 2, we review several specifications of spatial econometric models in a consistent framework of linear regression embedded with various spatial structures. The resulting model is in general nonlinear in parameters, and estimated by nonlinear maximum likelihood method. Section 3 summarizes the popular Moran-based statistics for test spatial correlation in the model. These tests include the classical Moran’s I index, the regular and robust version of two Lagrange multiplier based test statistics, one for a spatial lag model and one for a spatial error model. Section 4 introduces the methodology of bootstrap and develops the technique of spatial bootstrap for testing the spatially dependent data in a context of the spatial econometric model. In Section 5, two examples are used to illustrate the application of spatial bootstrap method. Finally, Section 6 concludes with hints on computing obstacle and challenge for future research.

2. Spatial Econometric Models

The basic idea of a spatial econometric model is to include spatial structure (spatial heterogeneity and spatial correlation among cross-section units of observations) in a classical linear regression model:

\[
y = X\beta + \varepsilon
\]

(1)

Where \(y\) is the dependent variable, \(X\) is a set of exogenous variables, and \(\varepsilon\) is the error term. As of spatial dependence, the model may be specified in two forms: a spatial lag

\(^1\) A test statistic is pivotal if its distribution does not depend on the unknown parameters under the null hypothesis.
model specifying the spatial correlation in the dependent variable, and a spatial error model allowing spatial autocorrelation in the nuisance error term. The former incorporates the first order autoregression of the neighboring units (countries, regions, or industries) as follows:

\[ y = \lambda Wy + X \beta + \epsilon \quad (2) \]

As in most of spatial econometric models (for example, Anselin [1988]), W is a row-standardized, zero-diagonal, spatial weight matrix of size NxN, where N is the number of units or observations. This is a very general description of the interconnected network structure among all units. For simplicity, we assume W is a time independent, location-based, binary-continuity, weight matrix, with entry’s value 1 for the adjacent neighbors and 0 otherwise. Accordingly, Wy is the weighted average of neighboring observations of y, and it is named the spatially lagged dependent variable. The coefficient \( \lambda \) is the corresponding spatial lag coefficient. We note that X may include spatially lagged (or weighted by W) exogenous variables.

On the other hand, a spatial error model relates the spatial dependence through the model errors. Due to omitted or missing variables and other unobservable factors, spatial interaction is modeled through the autocorrelated error structure. The error process can be further divided into spatial autoregressive error process and spatial moving-average error process. Let \( \epsilon \) be the regression error and the random variable u is i.i.d. With the first-order representation, we write the spatial error structure as: \( \epsilon = \rho W \epsilon + u \) and \( \epsilon = u - \theta Wu \) for the autoregressive and moving-average process, respectively. W\( \epsilon \) and Wu are the spatial weighted or lagged error term, \( \rho \) and \( \theta \) are the respective coefficients. Then the spatial autocorrelation error model is written as:

\[ y = X \beta + (I - \rho W)^{-1} u \quad (3) \]

Similarly, the spatial moving average error model is:

\[ y = X \beta + (I - \theta W)u \quad (4) \]

In summary, the above mentioned variations of spatial regression model can be combined into one representation:
\[ y^* = X^* \beta + \epsilon^* \quad (5) \]

If \( y^* = y, X^* = X, \epsilon^* = \epsilon \), this is the linear model (1) we begin with. If \( y^* = (I-\lambda W) y, X^* = X, \epsilon^* = \epsilon \), it is model (2). If \( y^* = (I-\rho W) y, X^* = (I-\rho W) X, \epsilon^* = u \), it is model (3). If \( y^* = (I-\theta W)^{-1} y, X^* = (I-\theta W)^{-1} X, \epsilon^* = u \), this is model (4). If \( y^* = (I-\theta W)^{-1}(I-\rho W) y, X^* = (I-\theta W)^{-1}(I-\rho W) X, \epsilon^* = u \), it is a combination of (3) and (4) or the spatial error ARMA model. Except for the basic model (1), the spatial model specifications are nonlinear in nature and the model parameters may not be identified without restrictions. We will consider the general case (5) only for the identified or estimable models.

Assuming i.i.d. error \( \epsilon^* \) with normal density of zero mean and constant covariance matrix \( \sigma^2 I \), the vector of parameter estimates \( (\hat{\beta}, \hat{\lambda}, \hat{\rho}, \hat{\theta}) \) is obtained by maximizing the scaled concentrated log-likelihood function for (5) as follows:

\[
\lll(\beta, \lambda, \rho, \theta \mid y, X, W) = -\frac{N}{2} \ln(\epsilon^* \epsilon^*) + \ln \left| \frac{\partial \epsilon^*}{\partial y} \right| = -\frac{N}{2} \ln \left[ (y^* - X^* \beta)(y^* - X^* \beta) \right] + \ln \left| \frac{\partial y^*}{\partial y} \right|
\]

The last term is the Jacobian transformation of the model specification under consideration. For example, \( \left| \frac{\partial y^*}{\partial y} \right| = |I - \lambda W| = \prod_{i=1}^{N} (1 - \lambda \omega_i) \) for (2) and

\[
\left| \frac{\partial y^*}{\partial y} \right| = \frac{1}{|I - \theta W|} = \frac{1}{\prod_{j=1}^{N} (1 - \theta \omega_j)} \]

for (4), where \( \omega_1, \omega_2, \ldots, \omega_N \) are the eigenvalues of the weight matrix \( W \). It is clear that for model stability, we require that

\[
\frac{1}{\omega_{\min}} < \lambda, \rho, \theta < \frac{1}{\omega_{\max}} = 1.
\]

Depending on the chosen specification, it is understood that not necessary all the parameters will be estimated.

Normality assumption is indeed too strong for the maximum likelihood estimation to be applicable for most of nonlinear regression models. For a spatial econometric model with potentially misspecified heteroscedastic error structure, quasi-maximum likelihood
(QML) method (White [1982]) is used so that the estimation is consistent and the inference is robust\(^2\).

### 3. Moran-Based Test Statistics

No matter which model we will be based upon for the final analysis, the chosen model must be correctly specified. Hypothesis testing is not only useful in selecting the best model but also essential for diagnostic checking the estimated model. Usually we begin with the basic linear regression model (1). Through pre-testing procedures, several model specifications emerge. Diagnostic testing based on the estimated model in the general form (5) is called for model validation. Diagnostic testing is different from the model pre-testing in that the post-test statistics depend on the estimated parameters and the model specification. The theoretical asymptotics for diagnostic checking is usually unknown, the test statistics designed for pre-testing purpose do not have the same asymptotic distribution when applied to diagnose the estimated model. In the case of spatial econometric models, there are additional issues of endogenous regressors, heteroscedasticity, and nonlinearity in the parameters.

The popular Moran-based tests available for spatial model diagnostics include the original Moran’s I statistic, two versions of LM-Error and LM-Lag tests based on Lagrange multiplier principle (see Anselin [1988], Florax, Folmer, and Rey [2003] for a good review). Moran’s I test statistic was originally developed to study the spatial correlation in random variables (Moran [1950]). Cliff and Ord [1973, 1981] applied this statistic to test the regression residuals for spatial dependence. Under the assumption of i.i.d. normal innovations, they derive the large sample distribution of the test statistic. Based on the general model (5), \( y^* = X^* \beta + \varepsilon^* \), we write the Moran’s I statistic as follows:

\[
I = \frac{\hat{\varepsilon}^* W \hat{\varepsilon}^*}{\hat{\varepsilon}^* \hat{\varepsilon}^*} \quad (6)
\]

Where \( \hat{\varepsilon}^* \) is the vector of regression residuals. Under the null hypothesis that the model is correctly specified and there is no spatial dependence in the model errors,

---

\(^2\) The asymptotic theory of QML for spatially autoregressive model has been developed by Lee [2004].
Moran’s I test statistic is asymptotically normally distributed with the mean $E(I)$ and variance $V(I)$ defined by:

$$E(I) = \frac{\text{trace}(MW)}{N - K}, \text{ where } M = I - X^* (X^{*\prime}X^*)^{-1} X^{*\prime}$$

$$V(I) = \frac{\text{trace}(MWMX) + \text{trace}((MW)^2)}{(N - K)(N - K + 2)} - E(I)^2$$

$N$ is the number of sample observations, and $K$ is the number of estimated parameters in the model. The standardized Moran’s I is conveniently used in practice.

Moran’s I is useful to test the spatial dependence in the regression, but it does not distinguish the two forms of spatial dependence in the model. Diagnostic tools for identifying a spatial autocorrelation in the error term or a spatial lag in the dependent variable are the LM-Error and LM-Lag tests. Define,

$$LM - \text{Error} = \left(\frac{\hat{\sigma}^2 W\hat{\epsilon}^2}{\hat{\epsilon}^2 \hat{\sigma}^2}\right)^2$$  \hspace{1cm} (7)$$

$$LM - \text{Lag} = \left(\frac{\hat{\sigma}^2 W\hat{y}^*}{\hat{\epsilon}^2 \hat{\sigma}^2}\right)^2 \frac{N}{J}$$  \hspace{1cm} (8)$$

where

$$T = \text{trace}(WW + W'W)$$

$$J = \frac{1}{N\hat{\sigma}^2} \left[(W\hat{y}^*)' M (W\hat{y}^*) + T\right]$$

$$\hat{y}^* = X^* \hat{\beta}, \hspace{0.5cm} \hat{\sigma}^2 = \hat{\epsilon}^2 \hat{\sigma}^2 / N$$

Under the assumption that the model error is independently normally distributed, LM-Error and LM-Lag test statistics are $\chi^2(1)$ distributed asymptotically (Burridge [1980], Anselin and Rey [1991]). There are robust versions of the LM tests, LM-Error* and LM-Lag*, which are designed to test for spatial error autocorrelation in the presence of a spatially lagged dependent variable and to test for endogenous spatial lag dependence in the presence of spatial error autocorrelation, respectively. Asymptotically, they are $\chi^2(1)$.
distributed as the classical versions (see Anselin, Bera, Florax, and Yoon [1996] for the derivation and simulation results).

\[
LM - Error^* = \left( \frac{\hat{\varepsilon}^* W \hat{\varepsilon}^* - T \frac{\hat{\varepsilon}^* W y^*}{\hat{\sigma}^2}}{\sigma^2 \sqrt{N J - T}} \right)^2
\]  
(9)

\[
LM - Lag^* = \left( \frac{\hat{\varepsilon}^* W y^* - \hat{\varepsilon}^* W \hat{\varepsilon}^*}{\sigma^2 \sqrt{N J - T}} \right)^2
\]  
(10)

The primary use of the above Moran-based tests is to pre-test the spatial independence in a linear regression model\(^3\). Anselin and Kelejian [1997] extend the test procedure to cover the cases with endogenous regressors including the spatially lagged dependent variable in the regression. Keleijian and Prucha [2001] consider the first-order spatially lagged dependent variable and spatial autocorrelation error in the null, and derive the asymptotic distribution for the Moran’s I test statistic. Their modification depends on the instrumental variable estimation and allows for limited dependent variable. Pinkse [2004] introduces a synthesis treatment of the Moran-flavored test statistics with nuisance parameters as in the case of a spatial probit model. The consensus is that limited distributions exist for Moran-based test statistics for spatial dependence under a set of general conditions for the i.i.d. normal model error structure.

Based on the linear regression model (1), Moran-based test statistics are routinely applied in the literature. Even if a regression model is corrected for spatial correlation, the estimated model in the general form (5) is subject to further scrutiny for the remaining spatial structure. Under the null hypothesis of no spatial dependence in a correctly specified and consistently estimated model, we can apply Moran-based test procedure for model validation. Consider the general model specification (5), the variables \(y^*\) and \(X^*\) are transformed in according to the spatial structure based on nonlinear maximum likelihood estimates of the model parameters. For diagnostic testing purpose, however, it is unlikely that the model error \(\varepsilon^*\) is i.i.d. normal. For a chosen spatial structure, the

\(^3\) Several strategies for model identification and selection using LM test statistics are discussed in Florax, Folmer, and Rey [2003].
model is nonlinear in the parameters and subject to misspecification. The point is that we do not have a well-known distribution for the diagnostic test statistics. This leads us to the empirical distribution approach such as bootstrapping. In the following, we introduce bootstrap method for diagnostic testing the spatial dependence in the data and model.

4. Bootstrap Method

In statistics, the bootstrap method relies on resampling from the observed data to approximate the probability distribution of the test statistics. Since its introduction by Efron [1979], the bootstrap has become a popular method for inference about the distribution of the parameters and related statistics. Without knowing the exact distributional properties of a test statistic or quantity, bootstrap has proven to be a practical method in many disciplines. See Efron and Tibshirani [1993], Davison and Hinkley [1997] for an introduction of the theory and applications of bootstrap method.

4.1. Introduction to Bootstrap

The bootstrap is based on the idea that the sample is a good representation of the underlying population distribution. It is a sort of Monte Carlo simulation in which we resample from the available sample information for inference about a parameter estimator or test statistic. Suppose $\tau$ is the test statistic of interest for a random variable $x$ with the probability distribution $F(x; \theta)$, where $\theta$ is the distribution parameter. We denote $\tau = \tau(F)$.

Let $\hat{F} = F(x; \hat{\theta})$ be the sample estimate of $F(x; \theta)$ and $\hat{\tau} = \tau(\hat{F})$, where $\hat{\theta}$ is the maximum likelihood estimator of $\theta$. Unless the asymptotic distribution of $\hat{\tau}$ is known, statistical inference is difficult to make. Even with the knowledge about the distribution $F(x; \theta)$, the asymptotic distribution of the test statistic $\hat{\tau}$ is complicated to derive and in general unknown. The alternative distribution-free bootstrap method is to resample $x$ with replacement, denoted $x^*$. Then we approximate the distribution $F(x; \theta)$ by $\hat{F}^* = F(x^*; \hat{\theta})$, from which we calculate the test statistic $\hat{\tau}^* = \tau(\hat{F}^*)$. After a large bootstrap sample of $\hat{\tau}^*$ is collected, statistical inference is made based on its empirical distribution instead of the asymptotic theory. By the law of large number, the empirical distribution of $\hat{\tau}^*$ approximates the theoretical distribution of $\hat{\tau}$ in large sample. That is, $\hat{\tau}^*$ converges to $\hat{\tau}$.
in distribution. The consistency property of $\hat{F}^* = F(x^*; \hat{\theta})$ which approximate $\hat{F} = F(x; \hat{\theta})$ is the foundation for bootstrap test to work well.

A standard bootstrap test proceeds as follows: First, based on the original N observations of the sample data we estimate $\theta$ for $F(x; \theta)$ and compute test statistic $\tau$. Secondly, we resample from the N observations of the original data series with replacement. Some observations may be selected more than once while some may not be picked at all. The result of the resampling is a bootstrap sample of size N. Then the bootstrap sample is used to estimate the model and to compute the test statistic exactly the same way as we did with the original sample in the beginning. Continue this process of replication until we collect sufficient number of bootstrap sample of the test statistic $\hat{\tau}$ so that its empirical distribution can be constructed and analyzed. For inference purpose, a bootstrap P-value defined by the fraction of the bootstrap sample that the bootstrap $\hat{\tau}^*$ greater than the estimated $\hat{\tau}$. Other tools such as confidence intervals may be bootstrapped as well for inference.

The performance of a bootstrap procedure depends on the empirical distribution in which the original data and model specification play the important role. Given the data resampling with replacement, it is understood that the results can not be generalized unless a large amount of replications is used. In addition, for bootstrap test, an asymptotically pivotal test statistic works better than the non-pivotal one.

4.2. Spatial Bootstrap Method

Techniques for bootstrap dependent data are the forefront activities of recent development (see Davison, Hinkley, and Young [2003], Horowitz [2003], MacKinnon [2002]). In time series, block bootstrap and sieve bootstrap are the popular examples. For spatial dependent data, earlier attempts were made since Cliff and Ord (1973) and Cressie (1980) but went unrecognized possibly due to computing constraints at that time. Recently there is no shortage of Monte Carlo methods in spatial econometric analysis, but the bootstrap method based on the estimated model and data is unseen except in a
different context (See Lee and Lahiri, 2002, for example). It is crucial that spatial structure must be preserved during data resampling. Because there is a fixed spatial weight matrix \( W \) of the full sample size preventing bootstrap by blocks or subsampling, it leaves only one option for us to bootstrap the regression residuals for a spatial econometric model. Under the assumption of independently and identically distributed residuals, for a fixed weight matrix \( W \) and a set of exogenous variables \( X \) we bootstrap the dependent variable \( Y \) through nonparametric resampling procedure as we perform for a classical linear regression model. Similar to the residual-based time series bootstrap, a recursive or reverse process through matrix inversion has to be used to bootstrap the dependent variable without destroying the spatial structure. The drawback of the residual-based bootstrapping is that the residuals must be i.i.d. by design. If it is normally distributed, then the parametric bootstrap is an alternative. We feel at the least the problem of heteroscedasticity due to cross-section data in the spatial model must be accounted for. Therefore, our bootstrap method for spatial econometric models is a hybrid of recursive wild bootstrap to be described in details below.

Suppose we observed a sample \((y, X)\) and a spatial weight matrix \( W \). For simplicity, we assume that the data matrix of exogenous variables \( X \) and the spatial weight matrix \( W \) are fixed by design. We further assume that the model error is independently and identically normally distributed, and the technique of quasi-maximum likelihood is used to obtain the consistent parameter estimators and their robust standard errors. Based on the estimated model (5), \( \hat{y}^* = X^* \hat{\beta} + \hat{\epsilon}^* \), we discuss the bootstrap method for Moran’s I statistic (6) \( I = \frac{\hat{\epsilon}^* W \hat{\epsilon}^*}{\hat{\epsilon}^* \hat{\epsilon}^*} \). For other models and other test statistics, the following illustration is the same.

From \( \hat{\epsilon}^* \), we start to bootstrap a sample of residuals \( \hat{\epsilon}^* \). If the normality assumption of the model error is acceptable, we could simply use the parametric bootstrap method. Let \( \hat{\epsilon}^* \) be a random draw from a normal distribution with zero mean and constant variance \( \hat{\sigma}^2 = \frac{\hat{\epsilon}^* \hat{\epsilon}^*}{N} \), and work around to get the bootstrap sample of \( y \). Because \( \hat{y}^* = X^* \hat{\beta} + \hat{\epsilon}^* \).

---

4 Pinkse and Slade (1998) propose a simulation-based test in probit models, in the same spirit of our bootstrap tests.
we apply a recursive or reverse process converting from $\hat{y}^*$ to $\tilde{y}$. For example, if $\hat{y}^* = (I - \hat{\lambda}W)y$ as in the case of a spatial lag model (2), then $\tilde{y} = (I - \hat{\lambda}W)^{-1}\hat{y}^*$. Given the fixed spatial weight matrix $W$, the bootstrap sample is $(\tilde{y}, X)$. Using this bootstrap sample, we repeat the estimation of model (5) and the computation of test statistic (6) to obtain $\tilde{y}^* = X^*\hat{\beta} + \tilde{e}^*$ and $I = \frac{\tilde{e}^*W\tilde{e}^*}{\tilde{e}^*\tilde{e}^*}$.

This idea of recursive bootstrap is similar to the time series autoregressive bootstrap, except that the recursion is essentially a reverse or matrix inverse process involving a full sample size spatial weight matrix $W$. The advantage is that it does not require pre-sample initialization as the time series bootstrap, but the disadvantage is the high computing cost of large matrix inversion. Although the spatial weight matrix $W$ may be sparse in most cases, its inverse in general is not. If the sample size is large, the repetitive computation of matrix inversion may prevent the practical application of spatial bootstrap. Indeed, if the sample size is large we still depend on the asymptotic theory for inference.

Parametric bootstrap is convenient and powerful when the model assumption is correct. Unfortunately, it is rare that the normality assumption of the model error will be satisfied. A nonparametric bootstrap is called for this purpose. From the estimated model, the bootstrap sample of the residuals $\tilde{e}^*$ is a reampling of $\hat{e}^*$ with replacement. Based on the general model (5), $\tilde{y}^* = X^*\hat{\beta} + \tilde{e}^*$, the regression model is nonlinear in parameters estimated with quasi-maximum likelihood method. To be consistent with the constant model variance, a re-scaled and re-centered residual is typically used in the bootstrap:

$$e^*_i = \sqrt{N} \left( \frac{\tilde{e}^*_i}{N-1} - \frac{1}{N} \sum_{j=1}^{N} \frac{\tilde{e}^*_j}{\sqrt{1-h^2_j}} \right), \quad i = 1, 2, ..., N$$

where $h$ is leverage or the diagonal vector of the hat matrix $X^*(X^*X^*)^{-1}X^*$. Further complication of a spatial econometric model using cross-section data is the presence of general heteroscedacity of unknown form. The method of wild bootstrap was invented to deal with this problem (Wu [1986]). That is, $\tilde{e}^*_i = e^*_i\nu_i$ with $\nu_i$ a random variable with mean 0 and variance 1. Several choices of $\nu_i$ are available in the literature (see Liu [1988],
Mammen [1993], Davidson and Flachaire [2001] among others. We adopt the most successful but the simplest form of two-point distribution as follows:

\[ \nu_i = \begin{cases} 
-1 & \text{with probability } 1/2 \\
1 & \text{with probability } 1/2 
\end{cases} \]

The rest works the same as the parametric method to get the bootstrap sample \((\tilde{y}, X)\) from \(\tilde{y}^* = X^*\hat{\beta} + \tilde{e}^*\) by matrix inversion. Finally, based on the bootstrap sample \((\tilde{y}, X)\), we estimate the model (5) \(\tilde{y}^* = X^*\hat{\beta} + \tilde{e}^*\) and compute the test statistic (6) \(\tilde{I} = \frac{\tilde{e}^* W \tilde{e}^*}{\tilde{e}^* \tilde{e}^*}\).

Other test statistics such as two versions of LM-Error and LM-Lag can be bootstrapped similarly.

It is known that in order to make bootstrap test work, a large number of the bootstrap sample should be obtained. The computing cost will be high because the model is estimated with nonlinear quasi-maximum likelihood method for each bootstrap sample. We replicate the bootstrap procedure for 999 times so that including the original one we generate 1000 bootstrap observations of the statistic. The empirical distribution function is built from these 1000 bootstrap statistics each with equal mass of 1/1000. The inference about the test statistics is to compute P-value as the fraction of the bootstrap test statistics which is greater than the original one. Taking the example of Moran’s I statistic, let \(I_0\) be the original statistic I obtained from the actual data and \(I_i\) through \(I_{999}\) are collected from the bootstrap \(\tilde{I}\) by resampling procedure described above. The resulting empirical distribution function is used to compare with the asymptotic normal distribution. By definition, bootstrap P-value of the Moran’s I test statistic is \(P(I_0) = \frac{\#\{I_i > I_0, i=1,2,\ldots,999\}}{1000}\), which may not be the same as that computed from the asymptotic distribution. If the bootstrap P-value is greater than the level of significance, says 5%, the estimated test statistic is to accept the null hypothesis based on the empirical distribution. We note that Moran’s I test statistic can be positive and negative in range, its P-value is the right-tail or left-tail probability depending on the value of the statistic. LM

---

5 The standardized normal variable \(\nu_i\) is an easy example, but in this case the distribution assumption is imposed for the general heteroscedasticity.

6 The notation \(#\{.\}\) is the number of observations satisfying the condition specified in \(\{.\}\).
5. Empirical Examples

In this section, we use two examples to demonstrate the application of spatial bootstrap procedure and provide empirical evidence of successfully bootstrap test statistics for spatial econometric models.

5.1. Example 1

By introducing the spatial econometric method to study urban crime incidents in Columbus Ohio, Anselin [1988] suggests a model that urban crime rate as a function of the family income and the real estates property value. Involving 49 neighborhoods in the study, it is found that the number of crime incidents reported is negatively related to the family income level and housing value in the neighborhood. The spatial dependence of the neighborhoods is an additional important factor for urban crime incidents. The basic regression model is:

\[
\text{(Crime Rate)} = \alpha + \beta \text{(Family Income)} + \gamma \text{(Housing Value)} + \epsilon
\]  
(E.1.1)

The estimated model is tested for spatial dependence in the regression residuals. Based on the asymptotic distribution assumptions of the test statistics involved--Moran’s I, regular and robust LM-Error, and LM-Lag, a spatially lagged dependent variable should be included in the regression as follows:

\[
\text{(Crime Rate)} = \alpha + \beta \text{(Family Income)} + \gamma \text{(Housing Value)} + \lambda \text{W (Crime Rate)} + \epsilon
\]  
(E.1.2)

where \( W \) is the simple first-order contiguity spatial weight matrix (row-standardized) for these 49 neighborhoods. In stead of relying on the asymptotic distribution assumptions for the test statistics, we apply nonparametric spatial bootstrap method to derive their empirical distributions and to compute P-values for the test statistics. First, we duplicate the results of Anselin [1988] including quasi-maximum likelihood estimates of the parameters, Moran’s I and LM test statistics. Then, in Table 1, we report our

---

7 All computations are performed using GPE2/GAUSS (see Lin [2001]) on a Pentium 4 2.4 Ghz Windows-XP system.
findings of spatial bootstrap, particularly P-values and confidence intervals for the test statistics.

From Table 1, by comparing the pre-test results, we find that both analytical and simulated Moran’s I test statistic are close and at 5% level of significance the test indicates the problem of spatial correlation. After the model has been corrected for spatial correlation with the addition of the first spatially lagged dependent variable in the regression equation (E.1.2), the post-test statistics confirm that there is no remaining spatial correlation in the residuals. In addition to Moran’s I, two versions of bootstrap LM tests validate the same conclusions from the asymptotic tests. In this case, we do not find discrepancy in the test results even though the asymptotic distribution is no longer a valid assumption for the estimated spatial lag model. The fact that both test results are similar is probably due to the theoretical distribution and empirical distribution being close enough for this particular data set and the estimated model.

[Insert Table 1]

5.2. Example 2

Using 28 provincial and city output data from 1978 to 2002, we study the regional economic convergence in China (Lin, Long, and Wu [2004]). Our contribution was to introduce the spatial econometric framework in an otherwise basic absolute convergence model:

\[
\text{Growth} = \alpha + \beta (\text{Initial Output}) + \epsilon \quad (E.2.1)
\]

Where (Initial Output) is measured by the provincial per capita GDP level in 1978, and Growth is the GDP growth rate over 25 years of development. From the vast volume of convergence literature, the estimated significant negative value of \( \beta \) indicates a convergence process in the economic development. It is now clear that through various stages of economic reform and policy change, the Chinese economy has developed with increasing dependence and cooperation across regions. Regional growth has seen to be influenced by the development of neighboring provinces and cities. In our previous study, we investigate the spatial effect in the growth and convergence. For simplicity, a first-order contiguity weight matrix \( W \) is used to specify the geographical connection of the
provinces and cities. In particular, the element of $W$ is 1 for the neighboring provinces and 0 otherwise. Furthermore, $W$ is row-standardized and zero-diagonal.

Reading from the second column of Table 2, based on the basic model (E.2.1), there are not much differences between the asymptotic and bootstrap test results. The conclusion is in general consistent and suggests that the original model suffers the ignorance of spatial correlation. Relying on the pre-test statistics for spatial correlation, we have found that the spatial autoregressive error model is a plausible model specification:

$$\text{Growth} = \alpha + \beta (\text{Initial Output}) + \varepsilon$$
$$\varepsilon = \rho \ W \varepsilon + u$$  \hspace{1cm} (E.2.2)

Based on the asymptotic theory, the post-test results for the estimated model do not reveal further evidence of the spatial correlation. P-values for all the test statistics are greater than 5% level of significance. Given the concern that we are working with a small sample of 28 heterogeneous cross-section observations and the spatial model is estimated by quasi-maximum likelihood method, the asymptotic distribution assumptions for the various test statistics are questionable. Here we bootstrap test statistics based on the estimated model (E.2.2) under the null hypothesis that there is no remaining spatial correlation. Although the estimated model (E.2.2) was reported in the previous study, the bootstrap post-test results shown in the third column of Table 2 indicate that the problem of spatial correlation is still in existence. The previously reported spatial model may be misspecified and subject to further scrutiny. The P-value of the bootstrap Moran’s I is too small (P=0.01), contradictory to the P-value of the theoretical counterpart (P=0.489). The alternative model specification is possibly a spatial lag model as follows:

$$\text{Growth} = \alpha + \beta (\text{Initial Output}) + \lambda \ W (\text{Growth}) + \varepsilon$$ \hspace{1cm} (E.2.3)

The estimated model and diagnostic test results are presented in the column 4 of Table 2. At 5% level of significance, all tests including the asymptotic and bootstrap variants are conclusive and consistent. Based on Maran’s I test, there is no remaining spatial correlation in the residuals. Based on various LM test statistics, the estimated model (E.2.3) is free of spatial dependence. One possible concern is the robust version of LM-Error test, which has a marginal P-value at 0.046. The revised model for the regional
convergence in China is better specified as a spatial lag model or a combined spatial lag and autoregressive error model. The later is not reported here due to space limitation. With a small sample and potentially a misspecified model, bootstrap demonstrates its power to distinguish a better candidate of the final model.

[Insert Table 2]

From these two examples, we learn that if the distribution assumption is known or correct, both asymptotic and bootstrap methods will yield the same and consistent conclusions. However, for diagnostic testing, it is rare that the estimated nonlinear model by quasi-maximum likelihood will satisfy the asymptotic distribution assumption. For post-test purpose, relying on the incorrect assumption of the distribution will in general lead to wrong conclusion about the model. Therefore it is often contradictory to the results based on bootstrap tests. The first example has a medium size of the sample, and the empirical distribution is in line with the theory. The results are consistent. The second example is more difficult to find a correct specification because we have only a small sample and therefore uncertain about the distribution assumption. Bootstrap method offers a better alternative instead of relying on the theoretical distribution. To make sure that we have a large number of bootstrap samples to work with, spatial econometric model can be tested and validated by the bootstrap methods.

6. Conclusion

This paper studies the bootstrap methods for diagnostic testing the spatial econometric models. A residual based recursive wild bootstrap is suggested to handle potential model misspecification due to nonlinearity in parameters and heteroscedasticity of unknown form. The methodology of spatial bootstrap has shown to be useful in identifying a better spatial structure than the classical approach which relies on the asymptotic distribution assumption. For a small sample problem, the bootstrap method is of particular important in diagnostic testing, provided a large number of bootstrap samples is used. Because of expensive computing cost associated with the nonlinear quasi-maximum likelihood estimation for each bootstrap spatial model, we are currently limited to examine only a few cases of small to medium size problems. Although the empirical results are promising, the formal asymptotic theory of spatial bootstrap is awaiting to be developed.
Theoretical and empirical investigation of the size and power of spatial bootstrap method is an important topic for future research.

References


Table 1. Crime Rate Regression (Anselin [1988])

<table>
<thead>
<tr>
<th>Model</th>
<th>E.1.1</th>
<th>E.1.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>α</td>
<td>68.619</td>
<td>45.079</td>
</tr>
<tr>
<td></td>
<td>(4.1015)</td>
<td>(6.4049)</td>
</tr>
<tr>
<td>β</td>
<td>-1.5973</td>
<td>-1.0316</td>
</tr>
<tr>
<td></td>
<td>(0.4464)</td>
<td>(0.42109)</td>
</tr>
<tr>
<td>γ</td>
<td>-0.27393</td>
<td>-0.26593</td>
</tr>
<tr>
<td></td>
<td>(0.15753)</td>
<td>(0.17309)</td>
</tr>
<tr>
<td>λ</td>
<td>0.43102</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.11067)</td>
<td></td>
</tr>
<tr>
<td>Log-likelihood</td>
<td>-187.38</td>
<td>-182.39</td>
</tr>
</tbody>
</table>

Moran’s I

<table>
<thead>
<tr>
<th></th>
<th>E.1.1</th>
<th>E.1.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computed</td>
<td>0.23564</td>
<td>0.037981</td>
</tr>
<tr>
<td></td>
<td>(P-Value = 0.001569)</td>
<td>(P-value = 0.21683)</td>
</tr>
<tr>
<td>Bootstrapped</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[5% 95%]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[2.5% 97.5%]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LM-Error</td>
<td>5.7230</td>
<td>0.14869</td>
</tr>
<tr>
<td></td>
<td>(P-value = 0.016744)</td>
<td>(P-value = 0.69979)</td>
</tr>
<tr>
<td>Bootstrapped</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[5% 95%]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[2.5% 97.5%]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LM-Lag</td>
<td>9.3634</td>
<td>0.013924</td>
</tr>
<tr>
<td></td>
<td>(P-value = 0.0022136)</td>
<td>(P-value = 0.90607)</td>
</tr>
<tr>
<td>Bootstrapped</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[5% 95%]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[2.5% 97.5%]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LM-Error*</td>
<td>0.79502</td>
<td>0.21634</td>
</tr>
<tr>
<td></td>
<td>(P-value = 0.77797)</td>
<td>(P-value = 0.64184)</td>
</tr>
<tr>
<td>Bootstrapped</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[5% 95%]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[2.5% 97.5%]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LM-Lag*</td>
<td>3.7199</td>
<td>0.081577</td>
</tr>
<tr>
<td></td>
<td>(P-value = 0.053768)</td>
<td>(P-value = 0.77517)</td>
</tr>
</tbody>
</table>

Note: For parameter estimates, the numbers in parentheses are the estimated standard errors. All bootstrap simulation of test statistics are based on 1000 bootstrap samples.
Table 2. Economic Convergence in China, 1978—2002 (Lin, Long, and Wu [2004])

<table>
<thead>
<tr>
<th>Model</th>
<th>E.2.1</th>
<th>E.2.2</th>
<th>E.2.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
<td>(\alpha)</td>
<td>(\beta)</td>
<td>(\rho)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(2.4160) (0.33221)</td>
<td>(-0.058628) (0.052548)</td>
<td>(0.43841) (0.16782)</td>
</tr>
<tr>
<td></td>
<td>(3.0197) (0.59245)</td>
<td>(-1.5943) (0.092431)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.9112) (0.31052)</td>
<td>(-1.0585) (0.061052)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(\rho_{\gamma})</td>
<td>(\lambda_{\gamma})</td>
<td></td>
</tr>
</tbody>
</table>

Note: For parameter estimates, the numbers in parentheses are the estimated standard errors. All bootstrap simulation of test statistics are based on 1000 bootstrap samples.