Efficient Estimation of the Semiparametric Spatial Autoregressive Model

Peter Robinson

THE INSTITUTE FOR FISCAL STUDIES
DEPARTMENT OF ECONOMICS, UCL
cemmap working paper CWP08/06

Efficient Estimation of the Semiparametric Spatial Autoregressive Model

P.M. Robinson*
Department of Economics, London School of Economics,
Houghton Street, London WC2A 2AE, UK

February 27, 2006

Abstract

Efficient semiparametric and parametric estimates are developed for a spatial autoregressive model, containing nonstochastic explanatory variables and innovations suspected to be non-normal. The main stress is on the case of distribution of unknown, nonparametric, form, where series nonparametric estimates of the score function are employed in adaptive estimates of parameters of interest. These estimates are as efficient as ones based on a correct form, in particular they are more efficient than pseudo-Gaussian maximum likelihood estimates at non-Gaussian distributions. Two different adaptive estimates are considered. One entails a stringent condition on the spatial weight matrix, and is suitable only when observations have substantially many "neighbours". The other adaptive estimate relaxes this requirement, at the expense of alternative conditions and possible computational expense. A Monte Carlo study of finite sample performance is included.

JEL Classifications: C13; C14; C21
Keywords: Spatial autoregression; Efficient estimation; Adaptive estimation; Simultaneity bias.

*Corresponding author: Tel. +44-20-7955-7516; fax: +44-20-7955-6592. *E-mail address:* p.m.robinson@lse.ac.uk.

1 Introduction

Spatial autoregressive models have proved a popular basis for statistical inference on spatial econometric data. Much of the spatial statistics literature has focussed on data recorded on a lattice, that is, it is regularly-spaced in two or more dimensions. This is an unlikely framework in economics, at best an approximation. Data recorded over geographical space are apt to be very irregularly spaced, such as at cities or towns, or aggregated across possibly contiguous regions, such as provinces or countries. A recent review of spatial econometrics is Arbia (2006). A statistical model that adequately describes dependence as a function of geographic distance is apt to be complicated, especially in the second kind of situation, and derivations of rules of large sample statistical inference under plausible conditions difficult; even for time series data, where there is a single dimension, inference in irregularly-spaced settings is not very well developed. On the other hand, cross-sectional correlation has been measured as a function of "economic distance", not necessarily in a geographic setting. Spatial autoregressive models are applicable in all these circumstances.

We wish to model an $n \times 1$ vector of observations $y = (y_1, ..., y_n)^T$, on a scalar variate y_i, T indicating transposition. We have an $n \times k$ matrix of constants $X = (x_1, ..., x_n)^T$, x_i being a $k \times 1$ vector, where $k \geq 1$. Let $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)^T$ be an $n \times 1$ vector of unobservable random variables, that are independent and identically distributed (iid) with zero mean and unit variance. Let l_n be the $n \times 1$ vector $(1, ..., 1)^T$. Finally, let W be a given $n \times n$ "weight" matrix, having zero diagonal elements and being row-normalized such that elements of each row sum to 1, so

$$Wl_n = l_n. \tag{1.1}$$

We assume that, for some scalars μ_0, σ_0 and λ_0, and some $k \times 1$ vector β_0,

$$y = \mu_0 l_n + \lambda_0 W y + X\beta_0 + \sigma_0 \varepsilon. \tag{1.2}$$

Here, μ_0 and $\sigma_0 > 0$ are unknown nuisance parameters, representing intercept and scale respectively: they can be estimated, but our focus is on the estimation of $\theta_0 = (\lambda_0, \beta_0^T)^T$, where $\lambda_0 \in [0, 1)$ and β_0 is non-null. It is taken for granted that there are no restrictions linking θ_0, μ_0 and σ_0. It is assumed that the matrix (l_n, X) has full column rank for sufficiently large n, and because $k \geq 1$ there must be at least one non-intercept regressor.

The weight matrix W has to be chosen by the practitioner. In view of the row-normalization, we can define it in terms of an underlying non-negative inverse "distance" measure d_{ij} such that W has (i, j)-th element

$$w_{ij} = \frac{d_{ij}}{\sum_{h=1}^{n} d_{ih}}. \tag{1.3}$$

However, the "distance" terminology is not taken to imply that W is necessarily a symmetric matrix.

In general, only large-sample statistical inference can be justified. Here, though we have mostly suppressed reference to the data size n for the sake of a concise notation, the row-normalization of W implies that as $n \to \infty$, y must be treated like a triangular array. In recent years considerable progress has been made in the econometric literature on developing asymptotic properties of various estimates for (1.2).

Ordinary least squares (OLS) comes first to mind. The OLS estimate of θ_0 in (1.2) (with μ_0 treated as unknown) is generally inconsistent, because, for each i, the i-th element of Wy is correlated with ε_i. This situation contrasts with the corresponding classical dynamic time series model in which the lagged dependent variable is uncorrelated with the disturbance. It mirrors the one identified by Whittle (1954), in case of multilateral autoregressive models on a lattice. He pointed out that the problem is corrected by using Gaussian maximum likelihood (ML) estimation: the determinant term in this, which OLS neglects, has non-negligible effect.

This is the case also in (1.2), and Lee (2004) has established desirable asymptotic properties of Gaussian ML here, namely $n^{\frac{1}{2}}$-consistency and asymptotic normality and efficiency. An alternative, if generally sub-optimal solution, is instrumental variables, and this has been justified by Kelejian and Prucha (1998, 1999), Lee (2003), Kelejian, Prucha and Yuzefovich (2003).

On the other hand, returning to OLS, Lee (2002) noticed that this can still be consistent, and even $n^{\frac{1}{2}}$-consistent and asymptotically normal and efficient under suitable conditions on W. In particular, he showed that consistency is possible if the d_{ij} in (1.3) are uniformly bounded and the $\sum_{j=1}^{n} d_{ij}$ tend to infinity with n, and $n^{\frac{1}{2}}$-consistent if the latter sums tend to infinity faster than $n^{\frac{1}{2}}$.

This can be simply illustrated in terms of a W employed in an empirical example of Case (1992), and stressed by Lee (2002). Data are recorded across p districts, in each of which are q farmers. Independence between farmers in different districts is assumed, and neighbours at each farm within a district are given equal weight. Due to (1.1) we have

$$W = I_p \otimes (q-1)^{-1} \left(l_q l_q^T - I_q \right). \tag{1.4}$$

In this setting, OLS is consistent if

$$q \to \infty, \quad \text{as } n \to \infty, \tag{1.5}$$

and $n^{\frac{1}{2}}$-consistent if

$$q/p \to \infty \quad \text{as } n \to \infty. \tag{1.6}$$

The procedure considered by Lee (2004), on the other hand, was actually interpreted not just as ML under Gaussianity, but also pseudo-ML under departures from Gaussianity, as is the case in many other settings. However, though $n^{\frac{1}{2}}$-consistency and asymptotic normality is relevant in the latter circumstances, asymptotic efficiency is not. When data-sets are not very large, precision is important, and since there is often reason not to take Gaussianity seriously, it is desirable to develop estimates which are efficient in non-Gaussian populations.

As is typically the case in time series models, building a non-Gaussian likelihood is most easily approached by introducing a non-normal distribution for the iid ε_i in (1.2), for example a Student-t distribution. Such a distribution may also involve unknown nuisance parameters, to be estimated alongside the original ones. We present limit distributional results for one-step Newton approximations to ML estimates in this case. However, there is rarely a strong reason for picking a particular parametric form for the underlying innovation density, and if this is mis-specified not only would the estimates not be asymptotically efficient (or necessarily more efficient than the Gaussian pseudo-ML estimates of Lee (2004)), but in some cases they may actually be inconsistent. As in other statistical problems, these drawbacks, as well as possible computational complications, do much to explain the popularity of Gaussian pseudo-ML procedures, and approximations to them.

On the other hand, the ability to "adapt" to distribution of unknown form is well-established in a number of other statistical and econometric models. Here the density f of ε_i in (1.2) is regarded as a nonparametric function, so that (1.2) is a semiparametric model, and f is estimated by smoothing. By a suitable implementation, it can then be possible to obtain estimates of the parameters of the model that are $n^{\frac{1}{2}}$-consistent and normal, and asymptotically as efficient as ones based on a correctly parameterized f. This was demonstrated by Stone (1975), for the simple location model with iid data, and then by Bickel (1982), Newey (1988) for regression models with iid errors, and by other authors in various other models. The main focus of the present paper is to develop such procedures for efficiently estimating the vector θ_0 in (1.2). The ability to adapt in (1.2) is not guaranteed. Our first result requires similar conditions on W to those that Lee (2002) imposed in justifying the $n^{\frac{1}{2}}$-consistency of OLS (i.e. (1.6) in case (1.4)). Our second result employs a bias-reduced estimate that, in case (1.4), requires only (1.5), though either W has also to be symmetric (as is the case in (1.4)) or ε_i has to be symmetrically distributed.

Our efficient estimates of θ_0 are described in the following section. Regularity conditions and theorem statements are presented in Section 3. Section 4 consists of a Monte Carlo study of finite sample behaviour. Proofs are left to an appendix.

2 Efficient Estimates

It is possible to write down an objective function that is a form of likelihood, employing a smoothed nonparametric estimate of the density f of the ε_i. However, not only is this liable to be computationally challenging to optimize, but derivation of asymptotic properties would be a lengthy business since, as is common in problems involving implicitly-defined extremum estimation, the proof of $n^{\frac{1}{2}}$-consistency and asymptotic normality has to be preceded by a consistency proof. The latter can be by-passed by the familiar routine of taking one Newton-type iterative step, based on the aforementioned "likelihood", from an initial

$n^{\frac{1}{2}}$-consistent estimate. This strategy is followed virtually uniformly in the adaptive estimation literature, and we follow it here.

It leads to the need to nonparametrically estimate not $f(s)$, but the score function

$$\psi(s) = -\frac{f'(s)}{f(s)}, \tag{2.1}$$

where throughout the paper the prime denotes differentiation. The bulk of work on adaptive estimation uses kernel estimates of f and f'. Kernel estimation is very familiar in econometrics, and can have important advantages. However, separate estimation of f and f' is necessary, and the resulting estimate of ψ is somewhat cumbersome.

More seriously, since f is liable to become small, use of an estimate in the denominator of (2.1) is liable to cause instability. It also causes technical difficulty, and typically some form of trimming is introduced. This requires introduction of a user-chosen trimming number, itself a disincentive to the practitioner. In addition, kernel-based adaptive estimates have, for technical reasons, featured sample-splitting (use of one part of the sample in the nonparametric estimation, and the other for the final parametric estimation) which is wasteful of data and introduces a further ambiguity, as well as the artificial device of discretizing the initial parameter estimate.

These drawbacks are overcome by employing a series estimate of ψ, as proposed by Beran (1976) in the context of estimation of the coefficients of a time series autoregressive model. Let $\phi_\ell(s)$, $\ell = 1, 2, ...$, be a sequence of smooth functions. For some user-chosen integer $L \geq 1$, define the vectors

$$\phi^{(L)}(s) = (\phi_1(s), ..., \phi_L(s))^T, \quad \bar{\phi}^{(L)}(s) = \phi^{(L)}(s) - E\left\{\phi^{(L)}(\varepsilon_i)\right\}. \tag{2.2}$$

Consider for $\psi(s)$ first the parametric form

$$\psi(s) = \bar{\phi}^{(L)}(s)^T a^{(L)}, \tag{2.3}$$

where $a^{(L)} = (a_1, ..., a_L)^T$ is a vector with unknown elements. The mean-correction in (2.2) imposes the restriction $E\{\psi(\varepsilon_i)\} = 0$. Under mild conditions on f, integration-by-parts allows $a^{(L)}$ to be identified by

$$a^{(L)} = \left[E\left\{\bar{\phi}^{(L)}(\varepsilon_i)\bar{\phi}^{(L)}(\varepsilon_i)^T\right\}\right]^{-1} E\left\{\bar{\phi}^{(L)\prime}(\varepsilon_i)\right\}. \tag{2.4}$$

Given a vector of observable proxies $\tilde{\varepsilon} = (\tilde{\varepsilon}_1, ..., \tilde{\varepsilon}_n)^T$, we approximate $a^{(L)}$ by $\tilde{a}^{(L)}(\tilde{\varepsilon})$, where, for a generic vector $q = (q_1, ..., q_n)^T$,

$$\tilde{a}^{(L)}(q) = W^{(L)}(q)^{-1} w^{(L)}(q), \tag{2.5}$$

with

$$W^{(L)}(q) = \frac{1}{n}\sum_{i=1}^{n} \Phi^{(L)}(q_i)\Phi^{(L)}(q_i)^T, \tag{2.6}$$

$$w^{(L)}(q) = \frac{1}{n}\sum_{i=1}^{n} \phi^{(L)\prime}(q_i), \tag{2.7}$$

and

$$\Phi^{(L)}(q_i) = \phi^{(L)}(q_i) - \frac{1}{n}\sum_{j=1}^{n}\phi^{(L)}(q_j). \tag{2.8}$$

Then defining

$$\psi^{(L)}\left(q_i; \tilde{a}^{(L)}(q)\right) = \Phi^{(L)}(q_i)^T \tilde{a}^{(L)}(q), \tag{2.9}$$

$\tilde{\psi}_{il} = \psi^{(L)}\left(\tilde{\varepsilon}_i; \tilde{a}^{(L)}(\tilde{\varepsilon})\right)$ is a proxy for $\psi(\varepsilon_i)$, and is inserted in the Newton step for estimating the unknown parameters.

However, Beran's (1976) asymptotic theory assumed that, for the chosen L, (2.3) correctly specifies $\psi(s)$. This amounts almost to a parametric assumption on f: indeed, if we took $L = 1$ and $\phi_1(s) = s$, $\psi(s)$ given by (2.3) is just the score function for the standard normal distribution. Newey (1988) considerably developed the theory by allowing L to go to infinity slowly with n, so that the right hand side of (2.3) is an approximation to (an infinite series representation of) $\psi(s)$, and thence (in regression with independent cross-sectional observations) establishing analogous adaptivity results to those, say, that Bickel (1982) had, using kernel estimation of ψ. Robinson (2005) developed Newey's (1988) asymptotic theory further, in the context of stationary and nonstationary fractional time series models. In the asymptotic theory, L can be regarded as a smoothing number analogous to that used in a kernel approach, but no other user-chosen numbers or arbitrary constructions are required in the series approach, where indeed some regularity conditions are a little weaker than those in the kernel-based literature.

We follow the series estimation approach here, and for ease of reference mainly follow the notation of Robinson (2005). To provide further details for the adaptive estimation of our θ_0, consider the $n \times 1$ vector

$$e(\theta) = (e_1(\theta), ..., e_n(\theta))^T = (I - \lambda W)y - X\beta, \tag{2.10}$$

for $\theta = \left(\lambda, \beta^T\right)^T$, and any scalar λ and $k \times 1$ vector β. From (1.2)

$$\sigma_0 \varepsilon = e(\theta_0) - E\left\{e(\theta_0)\right\}. \tag{2.11}$$

Accordingly, given an initial estimate $\tilde{\theta}$ of θ_0, consider as a proxy for the vector $\sigma_0\varepsilon$ the vector $E(\tilde{\theta})$, where

$$E(\theta) = e(\theta) - l_n \frac{1}{n}\sum_{i=1}^{n} e_i(\theta). \tag{2.12}$$

We can estimate σ_0^2 by $\tilde{\sigma}^2 = \tilde{\sigma}^2(\tilde{\theta})$, where

$$\tilde{\sigma}^2(\theta) = \frac{1}{n} E(\theta)^T E(\theta). \tag{2.13}$$

Thus our proxy $\tilde{\varepsilon}$ for ε is given by

$$\tilde{\varepsilon} = E(\tilde{\theta})/\tilde{\sigma}. \tag{2.14}$$

We find it convenient to write

$$\tilde{\psi}_{iL} = \tilde{\psi}_{iL}(\tilde{\theta}, \tilde{\sigma}), \tag{2.15}$$

where

$$\tilde{\psi}_{iL}(\theta, \sigma) = \Phi^{(L)}\left(E_i(\theta)/\sigma\right)^T \tilde{a}^{(L)}\left(E(\theta)/\sigma\right). \tag{2.16}$$

Now introduce the $n \times (k+1)$ matrix of derivatives

$$e' = \left[\frac{\partial e(\theta)}{\partial \lambda}, \ \frac{\partial e(\theta)}{\partial \beta^T}\right]^T, \tag{2.17}$$

in which

$$\frac{\partial e(\theta)}{\partial \lambda} = -Wy, \tag{2.18}$$

$$\frac{\partial e(\theta)}{\partial \beta} = -X^T, \tag{2.19}$$

for all θ. With e_i' denoting the i-th column of e' write

$$E_i' = e_i' - \frac{1}{n}\sum_{j=1}^{n} e_j'. \tag{2.20}$$

Now define

$$R = \sum_{i=1}^{n} E_i' E_i'^T, \tag{2.21}$$

and

$$r_L(\theta, \sigma) = \sum_{i=1}^{n} \tilde{\psi}_{iL}(\theta, \sigma) E_i', \tag{2.22}$$

and let

$$\tilde{\mathcal{I}}_L(\theta, \sigma) = \frac{1}{n}\sum_{i=1}^{n} \tilde{\psi}_{iL}^2(\theta, \sigma), \tag{2.23}$$

so $\tilde{\mathcal{I}}_L(\tilde{\theta}, \tilde{\sigma})$ estimates the information measure

$$\mathcal{I} = E\psi(\varepsilon_i)^2. \tag{2.24}$$

Our first adaptive estimate of θ_0 is

$$\hat{\theta}_A = \tilde{\theta} - \left[\tilde{\mathcal{I}}_L(\tilde{\theta}, \tilde{\sigma})R\right]^{-1} r_L(\tilde{\theta}, \tilde{\sigma}). \tag{2.25}$$

(There is a typographical error in the corresponding formula (2.2) of Robinson (2005): " + " should be " − ".) Define

$$s_L(\theta, \sigma) = r_L(\theta, \sigma) + \left[\begin{array}{c} tr\left\{W(I_n - \lambda W)^{-1}\right\} \\ 0 \end{array}\right], \tag{2.26}$$

so s_L and r_L differ only in their first element. Our second adaptive estimate of θ_0 is

$$\hat{\theta}_B = \hat{\theta} - \left[\tilde{\mathcal{I}}_L(\tilde{\theta}, \tilde{\sigma}) R \right]^{-1} s_L(\tilde{\theta}, \tilde{\sigma}). \tag{2.27}$$

There are some practical issues outstanding. One is the choice of the functions $\phi_\ell(s)$. As in Newey (1988), Robinson (2005), we restrict to "polynomial" forms

$$\phi_\ell(s) = \phi(s)^\ell, \tag{2.28}$$

for some chosen function $\phi(s)$. For example,

$$\phi(s) = s, \tag{2.29}$$

$$\phi(s) = \frac{s}{(1+s^2)^{\frac{1}{2}}}, \tag{2.30}$$

where the boundedness in (2.30) can help to reduce other technical assumptions. Next, the choice of L is discussed in some detail by Robinson (2005); asymptotic theory provides little guidance here, indeed it delivers an upper bound on the rate at which L can increase with n, but no lower bound. Since the upper bound is only logarithmic in n, it seems that values of L should be used that are far smaller than n. Discussion of the choice of $\tilde{\theta}$ is postponed to the following section.

For completeness, we also consider the fully parametric case, where $f(s;\tau_0)$ is a prescribed parametric form for $f(s)$, with τ_0 an unknown $m \times 1$ vector, on the basis of which define $\widehat{\tau} = \arg\min_{\mathcal{T}} \sum_i \log f(E_i(\widetilde{\theta})/\widetilde{\sigma};\tau)$ for a subset $\mathcal{T}$ of $\mathbb{R}^m$, and, with $\psi(s;\tau) = (\partial/\partial s) f(s;\tau)/f(s;\tau)$

$$\tilde{\mathcal{I}}_L(\theta, \sigma, \tau) = n^{-1} \sum_i \psi\left(E_i(\theta)/\sigma; \tau\right)^2, \tag{2.31}$$

$$r_L(\theta, \sigma, \tau) = \sum_i \psi\left(E_i(\theta)/\sigma; \tau\right) E_i'(\theta). \tag{2.32}$$

Define also

$$s_L(\theta, \sigma, \tau) = r_L(\theta, \sigma, \tau) + \begin{bmatrix} tr\left\{W(I_n - \lambda W)^{-1}\right\} \\ 0 \end{bmatrix}. \tag{2.33}$$

Our two parametric estimates are

$$\hat{\theta}_C = \tilde{\theta} - \left[\tilde{\mathcal{I}}_L(\tilde{\theta}, \tilde{\sigma}, \widehat{\tau}) R \right]^{-1} r_L(\tilde{\theta}, \tilde{\sigma}, \widehat{\tau}), \tag{2.34}$$

$$\hat{\theta}_D = \tilde{\theta} - \left[\tilde{\mathcal{I}}_L(\tilde{\theta}, \tilde{\sigma}, \widehat{\tau}) R \right]^{-1} s_L(\tilde{\theta}, \tilde{\sigma}, \widehat{\tau}), \tag{2.35}$$

the second being a "bias-corrected" version of the first.

3 Asymptotic Normality and Efficiency

We introduce first the following regularity conditions

Assumption 1: *For all sufficiently large n, the weight matrix W has non-negative elements that are uniformly of order $O(1/h_n)$, where*

$$h_n/n^{\frac{1}{2}} \to \infty, \quad \text{as } n \to \infty, \tag{3.1}$$

and has zero diagonal, satisfies (1.1), and is such that the elements of $l_n^T W$ and $l_n^T S^{-1}$ are uniformly bounded, where $S = I_n - \lambda_0 W$.

Assumption 2: *The elements of the x_i are uniformly bounded constants, and the matrix*

$$\Omega = \lim_{n\to\infty} \frac{1}{n} \begin{bmatrix} (GX\beta_0)^T \\ X^T \end{bmatrix} [GX\beta_0, X] \tag{3.2}$$

exists and is positive definite, where $G = WS^{-1}$.

Assumption 3: *The ε_i are iid with zero mean and unit variance, and probability density function $f(s)$ that is differentiable, and*

$$0 < \mathcal{I} < \infty. \tag{3.3}$$

Assumption 4: *$\phi_\ell(s)$ satisfies (2.28), where $\phi(s)$ is strictly increasing and thrice differentiable and is such that, for some $\kappa \geq 0$, $K < \infty$,*

$$|\phi(s)| \leq 1 + |s|^K \tag{3.4}$$

$$|\phi'(s)| + |\phi''(s)| + |\phi'''(s)| \leq C\left(1 + |\phi(s)|^K\right), \tag{3.5}$$

where C is throughout a generic positive constant.

Denote by $\eta = 1 + 2^{\frac{1}{2}} \simeq 2.414$ and $\varphi = (1 + |\phi(s_1)| / \{\phi(s_2) - \phi(s_1)\}$, $[s_1, s_2]$ being an interval on which $f(s)$ is bounded away from zero.

Assumption 5:

$$L \to \infty, \quad \text{as } n \to \infty, \tag{3.6}$$

and either

(i) $\kappa = 0$, $E\varepsilon_i^4 < \infty$, and

$$\lim_{n\to\infty} \left(\frac{\log n}{L}\right) > 8\{\log \eta + \max(\log \varphi, 0)\} \simeq 7.05 + 8\max(\log \varphi, 0), \tag{3.7}$$

or (ii) $\kappa > 0$ for some $\omega > 0$ the moment generating function $E\left(e^{t|\varepsilon_i|^\omega}\right)$ exists for some $t > 0$, and

$$\liminf_{n\to\infty} \left(\frac{\log n}{L \log L}\right) > \max\left(\frac{8K}{\omega}, \frac{4\kappa(\omega+1)}{\omega}\right), \tag{3.8}$$

or (iii) $\kappa > 0$, ε_i is almost surely bounded, and

$$\liminf_{n\to\infty}\left(\frac{\log n}{L\log L}\right) > 4\kappa. \tag{3.9}$$

Assumption 6: *As $n \to \infty$*

$$\tilde{\theta} - \theta_0 = O_p(n^{-\frac{1}{2}}), \quad \tilde{\sigma}^2 - \sigma_0^2 = O_p(n^{-\frac{1}{2}}). \tag{3.10}$$

The proof of the following theorem is left to the Appendix.

Theorem A *Let (1.2) hold with $\lambda_0 \in [0,1)$, and let Assumptions 1-6 hold. Then as $n \to \infty$*

$$n^{\frac{1}{2}}\left(\hat{\theta}_A - \theta_0\right) \to_d N\left(0, \frac{\sigma_0^2}{\mathcal{I}}\Omega^{-1}\right), \tag{3.11}$$

where the limit variance matrix is consistently estimated by $\left(\tilde{\sigma}^2/\tilde{\mathcal{I}}_L(\tilde{\theta},\tilde{\sigma})\right) nR^{-1}$.

Remark 1 For the Gaussian pseudo-ML estimate, Lee (2004) finds the limiting variance matrix to be $\sigma^2\Omega^{-1}$. Since $\mathcal{I} \geq 1$, $\hat{\theta}_A$ achieves an efficiency improvement over this when ε_i is non-Gaussian.

Remark 2 Various initial estimates that satisfy Assumption 1 are available in the literature. This is the case under (3.1) if $\tilde{\theta}$ is the OLS estimate of θ_0 (see Lee, 2002). Other possibilities are the Gaussian pseudo-ML estimate, and various IV estimates.

Remark 3 In view of Assumption 2, β_0 cannot be the null vector (cf. Lee (2004)), so Theorem A cannot be used to test $\beta_0 = 0$, though it can be used to test exclusion of a proper subset of the elements of x_i. It can also be used to test the hypothesis of no spatial dependence, $\lambda_0 = 0$, and in this case the limit distribution in the Theorem is true even if (3.1) does not hold, indeed h_n can be regarded as fixed with respect to n, and so designs with only very few "neighbours" are covered. For non-Gaussian data, the tests provided by the Theorem are in general expected to be locally more powerful than ones based on existing estimates.

Remark 4 Assumption 1 is discussed by Lee (2002). In case W is given by (1.4), $h_n \sim q$, so condition (3.1) is equivalent to (1.6), and the rest of Assumption 1 is satisfied.

Remark 5 Assumptions 3-5 are essentially taken from Robinson (2005), where they are discussed. The main implications are that if we choose bounded $\phi(s)$ then a fourth moment condition on ε_i suffices, with a relatively mild upper bound restriction on the rate of increase of L (see (i)). For unbounded $\phi(s)$, we have a choice between moment generating function (ii) and boundedness (iii) requirements on ε_i, where the condition on L is weaker in the latter case, but still stronger than that of (i) of Assumption 5.

Remark 6 It would be possible to obtain analogous results for a non-linear regression extension of (1.2), in which the elements of $X\beta_0$ are replaced by the nonlinear-in-β_0 functions $g(x_i;\beta_0)$, $i = 1, ..., n$, where g is a smooth function of known form. With respect to the initial estimate $\tilde{\theta}$ it would seem that non-linear least squares can be shown to satisfy Assumption 6 under (3.1), by extending the arguments of Lee (2002).

Remark 7 In practice further iteration of (2.25) may be desirable. This would not change the limit distribution, but can improve higher-order efficiency (Robinson, 1988).

By far the most potentially restrictive of the conditions underlying Theorem 1 is (3.1) of Assumption 1. It is never really possible to gauge the relevance of an asymptotic condition such as this to a given, finite, data set. However, if, in the simple case where W is given by (1.4), q is small relative to p, one expects that $\hat{\theta}_A$ may be seriously biased, and the normal approximation poor; the same will be true of OLS.

Results of Lee (2004) on the Gaussian pseudo-MLE hint at how it may be possible to relax (3.1). To best see this we temporarily modify the model (1.2) to

$$y = \lambda_0 Wy + Z\gamma_0 + \sigma_0 \varepsilon. \tag{3.12}$$

If an intercept is allowed, as in (1.2), then l_n is a column of Z, $Z = (l_n, X)$, and $\gamma_0 = (\mu_0, \beta_0^T)^T$. But it is also possible that no intercept is allowed, unlike in (1.2), in which case $Z = X$ and $\gamma_0 = \beta_0$ (and $\mu_0 = 0$). The form (3.12) is the most usual in the literature. Lee (2004) shows the Gaussian pseudo-MLE $\hat{\xi}^* = \left(\hat{\lambda}^*, \hat{\gamma}^{*T}, \hat{\sigma}^{*2}\right)^T$ of $\xi_0 = \left(\lambda_0, \gamma_0^T, \sigma_0^2\right)^T$ is $n^{\frac{1}{2}}$-consistent and asymptotically normal, under mild conditions that do not even require that h_n diverge (i.e. in (1.4), q can remain fixed). However, even under Gaussianity, $\hat{\lambda}^*$ and $\hat{\sigma}^{*2}$ are independent in the limit distribution if h_n does not diverge, suggesting that adaptive estimation of λ_0, γ_0 is not possible in this scenario. Lee (2004) finds, however, that the limit covariance matrix of $\hat{\xi}^*$ simplifies when

$$h_n \to \infty, \quad \text{as } n \to \infty, \tag{3.13}$$

(i.e. (1.5) under (1.4)). His formulae indicate that $\left(\hat{\lambda}^*, \hat{\gamma}^{*T}\right)^T$ and $\hat{\sigma}^{*2}$ will then be asymptotically independent if $E\left(\varepsilon_i^3\right) l_n^T GZ\gamma_0/n \to 0$, $E\left(\varepsilon_i^3\right) l_n^T/n \to 0$, as $n \to \infty$. This is true if ε_i is normally distributed, and somewhat more generally, e.g. if ε_i is symmetrically distributed. Reverting now to our model (1.2) and with $(\hat{\lambda}^*, \hat{\beta}^{*T}, \hat{\sigma}^{*2})$ denoting the Gaussian pseudo-MLE of $\left(\lambda_0, \beta_0^T, \sigma_0^2\right)$, analogously $\left(\hat{\lambda}^*, \hat{\beta}^{*T}\right)$ and $\hat{\sigma}^{*2}$ are asymptotically independent if

$$E(\varepsilon_i^3) l_n^T GHX\beta_0/n \to 0, \quad E(\varepsilon_i^3) l_n^T HX/n \to 0, \text{ as } n \to \infty, \tag{3.14}$$

where $H = I_n - l_n l_n^T/n$. The latter limit always holds (since $l_n^T H = 0$), indeed the left hand side is the null vector for all n. The first limit holds if W is symmetric (because (1.1) then implies $l_n^T W = l_n^T$), and again the left hand side is zero for all n. (Such symmetry obtains in (1.4).) Thus if we focus on λ_0 and slope parameters only, their estimates are independent of $\hat{\sigma}^{*2}$ more generally than Lee (2004) claims, to enhance further the value of his results.

These observations suggest that if we start from a full, but possibly non-Gaussian likelihood, in the spirit of Whittle (1954), so that it contains a Jacobian factor $\det^{\frac{1}{2}}\{I_n - \lambda W\}$, we will both achieve sufficient bias-correction to enable relaxation of (3.1) to (3.13), and the information matrix block-diagonality necessary for adaptivity, so long as either W is symmetric or ε_i is symmetrically distributed (the moments $E(\varepsilon_i^3)$ in the above discussion are replaced by $E\left(\varepsilon_i \psi(\varepsilon_i)^2\right)$.

The proof of the following theorem is omitted, due to the preceding discussion, and the fact that our proof of Theorem A focusses on how the bias problem is resolved by (3.1), and not only is this aspect now unnecessary, but the remainder of the proof details are partly covered by some given in the proof of Theorem A, while others are relatively straightforward.

Theorem B *Let (1.2) hold with $\lambda_0 \in [0,1)$, and let Assumptions 1-6 hold with (3.1) relaxed to (3.13), and let either W be symmetric or ε_i be symmetrically distributed. Then as $n \to \infty$,*

$$n^{\frac{1}{2}}\left(\widehat{\theta}_B - s\right) \to_d N\left(0, \frac{\sigma_0^2}{\mathcal{I}}\Omega^{-1}\right), \tag{3.15}$$

where the limit variance matrix is consistently estimated by $\left(\tilde{\sigma}^2 / \tilde{\mathcal{I}}_L(\tilde{\theta}, \tilde{\sigma})\right) nR^{-1}$.

Remark 8 In general $\hat{\theta}_B$ can be expensive to compute because the second component of $s_L(\theta, \sigma)$ involves the inverse of an $n \times n$ matrix. However, in some special cases it is very simple, e.g. in case W is given by (1.4), we have

$$tr\left\{W(I_n - \lambda W)^{-1}\right\} = \frac{n\lambda}{(q-1+\lambda)(1-\lambda)}. \tag{3.16}$$

Remark 9 We cannot use OLS for $\tilde{\theta}$, $\tilde{\sigma}^2$ if (3.1) does not hold. We can, however, use an IV estimate, such as those of Kelejian and Prucha (1998, 1999), Lee (2003), or the Gaussian pseudo-MLE of Lee (2004).

Remark 10 As in other models, under symmetry of ε_i it is also possible to adapt with respect to the estimation of μ_0 in (1.2).

With respect to $\hat{\theta}_C$ and $\hat{\theta}_D$ we introduce the following additional assumptions.

Assumption A7 *$\mathcal{T}$ is compact and τ is an interior point of $\mathcal{T}$.*

Assumption A8 *For all $\tau \in \mathcal{T} - \{\tau_0\}$, $f(s;\tau) \neq f(s;\tau_0)$ on a set of positive measure.*

Assumption A9 *In a neighbourhood $\mathcal{N}$ of τ_0, $\log f(s;\tau)$ is thrice continuously differentiable in τ for all s and*

$$\int_{-\infty}^{\infty}\left\{\sup_{\mathcal{N}}\left|f^{(k)}(s;\tau)\right|+\sup_{\mathcal{N}}\left|f^{(k,\ell)}(s;\tau)\right|+\sup_{\mathcal{N}}\left|f^{(k,\ell,m)}(s;\tau)\right|\right\}ds<\infty, \quad (3.17)$$

where $f^{(k)}$, $f^{(k,\ell)}$, $f^{(k,\ell,m)}$ represent partial derivatives of f with respect to the k-th, the k-th and ℓ-th, and the k-th, ℓ-th and m-th elements of τ, respectively.

Assumption A10 $\Psi = E\left((\partial/\partial\tau)\log f(\varepsilon_i;\tau_0)(\partial/\partial\tau^T)\log f(\varepsilon_i;\tau_0)\right)$ *is positive definite.*

Theorem C *Let (1.2) hold with* $\lambda_0 \in [0,1)$, and l*et Assumptions 1-3 and 6-10 hold. Then as $n\to\infty$, $n^{\frac{1}{2}}(\hat{\theta}_C-\theta_0)$and $n^{\frac{1}{2}}(\widehat{\tau}-\tau_0)$ converge to independent $N(0,(\sigma_0^2/\mathcal{I})\Omega^{-1})$, and $N(0,\Psi^{-1})$ vectors respectively, where the limiting covariance matrices are consistently estimated by $\left(\tilde{\sigma}^2/\tilde{\mathcal{I}}_L(\tilde{\theta},\tilde{\sigma},\widehat{\tau})\right)nR^{-1}$ and*

$$\left\{n^{-1}\sum_{t=1}^{n}\left[(\partial/\partial\tau)\log f\left(E_i(\tilde{\theta})/\tilde{\sigma};\widehat{\tau}\right)\right]\left[(\partial/\partial\tau^T)\log f\left(E_i(\tilde{\theta})/\tilde{\sigma}_2;\widehat{\tau}\right)\right]\right\}^{-1}, \quad (3.18)$$

respectively.

Theorem D *Let (1.2) hold with* $\lambda_0 \in [0,1)$, and *let Assumptions 1-3 and 6-10 hold with (3.1) relaxed to (3.13), and let either W be symmetric or ε_i* be symmetrically distributed. *Then as $n\to\infty$, $n^{\frac{1}{2}}(\hat{\theta}_D-\theta_0)$and $n^{\frac{1}{2}}(\widehat{\tau}-\tau_0)$ converge to independent $N(0,(\sigma_0^2/\mathcal{I})\Omega^{-1})$, and $N(0,\Psi^{-1})$ vectors respectively, where the limiting covariance matrices are consistently estimated by $\left(\tilde{\sigma}^2/\tilde{\mathcal{I}}_L(\tilde{\theta},\tilde{\sigma},\widehat{\tau})\right)nR^{-1}$ and (3.18) respectively.*

The proofs would require first an initial consistency proof for the implicitly-defined extremum estimate $\widehat{\tau}$, and are omitted because they combine elements of the proof of Theorem A with relatively standard arguments.

Remark 11 The Gaussian MLE can in general be expensive to compute due to the determinant factor, as discussed by Kelejian and Prucha (1999). However, the limit distribution of this estimate is the same as that of $\hat{\theta}_C$ and $\hat{\theta}_D$ when these are based on $f(s;\tau)=(2\pi)^{-1/2}\exp(-s^2/2)$, $\psi(s;\tau)=s$. Indeed such $\hat{\theta}_D$ represents a Newton step to the Gaussian MLE.

4 Finite Sample Performance

The behaviour of our adaptive estimates in finite sample sizes was examined in a small Monte Carlo study. The spatial weight matrix W given by (1.4) was employed, with three different choices of (p, q): (8,12), (11,18), (14,28). These correspond to $n = 96$, 198 and 392, and are intended to represent a slow approach to the asymptotic behaviour of (3.1). For each n, scalar explanatory variables $x_1, ..., x_n$ were generated as iid uniform $(0, 1)$ observations, and then kept fixed throughout the study, to conform to the non-stochastic aspect of Assumption 2. The ε_i were generated from each of the following 5 distributions.

(a) Normal, $\varepsilon_i \sim N(0, 1)$,

(b) Bimodal Mixture normal, $\varepsilon_i = u/\sqrt{10}$ where $pdf(u) = \frac{.5}{\sqrt{2\pi}} \exp\left(-\frac{(u-3)^2}{2}\right) + \frac{.5}{\sqrt{2\pi}} \exp\left(-\frac{(u+3)^2}{2}\right)$,

(c) Unimodal Mixture normal, $\varepsilon_i = u/\sqrt{2.2}$, where $pdf(u) = \frac{.05}{\sqrt{50\pi}} \exp\left(-\frac{u^2}{50}\right) + \frac{.95}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$,

(d) Laplace, $f(s) = \exp\left(-|s|\sqrt{2}\right)\sqrt{2}$,

(e) Student t_5, $\varepsilon_i = u\sqrt{3/5}$, where $u \sim t_5$.

These are fairly standard choices in Monte Carlo studies of adaptive estimates. All of them are scaled to have variance 1, as in Assumption 3. Case (a) has finite moments of degree 4 only.

On each of 1000 replications, y was generated from (1.2) with $\mu_0 = 0$, $\gamma_0 = 1$, $\beta_0 = 1$, and with two different λ_0 values, 0.4 and 0.8, for each of the 3 n values and 5 ε_i distributions. Both $\hat{\theta}_A$ and $\hat{\theta}_B$ were computed in each case, for both choices (2.29) and (2.30) of $\phi(s)$ (respectively denoted "1" and "2" in the tables below), and for $L = 1, 2, 4$. We took $\tilde{\theta}$ to be OLS.

Lee (2004) featured the design (1.4) in his Monte Carlo study of the Gaussian MLE. The two experiments are not closely comparable. He employed a wider range of n and x_i, while restricting to Gaussian ε_i and a single λ_0 (0.5), and with no comparison with other estimates. Our study looks at relative efficiency over a range of distributions, our examination of two values of λ_0 turns out to throw light on the bias issue, and we explore aspects of implementation which do not arise for his estimate. Nevertheless. we shall have occasion to refer to his results, and make some remarks about computational issues prompted by both studies.

Monte Carlo bias, variance and mean squared error (MSE) were computed in each of the $2 \times 2 \times 2 \times 3 \times 3 \times 5 = 360$ cases. In view of the potential impact of bias, Tables 1 and 2 report Monte Carlo bias of both elements, $\widetilde{\lambda}$, $\widetilde{\beta}$, of the initial estimate, OLS. For $\lambda_0 = 0.4$ the bias of $\widetilde{\lambda}$ actually increases with n, suggesting that a faster increase of q/p would give better results here. For

$\lambda_0 = 0.8$, biases for the smaller n are greater, they fall then rise with a slight net reduction. We will encounter some consequences of this bias on $\hat{\theta}_A$ and $\hat{\theta}_B$. The bias of $\widetilde{\beta}$ is much smaller, a phenomenon found also for $\hat{\theta}_A$ and $\hat{\theta}_B$, and by Lee (2004) for his estimate.

(Tables 1 and 2 about here)

Each of Tables 3-18 presents one of the rival efficiency measures, relative variance and relative MSE, comparing one element of $\hat{\theta}_A = (\widehat{\lambda}_A, \widehat{\beta}_A)^T$ with OLS in one ε_i distribution, and covers all n, λ_0, L and ϕ. To conserve on space we have omitted the normal distribution (a) results. Here, one expects $\hat{\theta}_A$ to be worse than OLS for all $L \geq 1$ when $\phi = (2.30)$, and to deteriorate with increaing L when $\phi = (2.29)$. This turned out to be the case, but though the ratios peaked at 1.4447 (in case of relative variance of $\widehat{\lambda}_A$ for $\lambda_0 = 0.8$, $n = 96$, $L = 2$, $\phi = (2.30)$), they were mostly less than 1.1.

(Tables 3-6 about here)

Tables 3-6 concern the bimodal mixture normal (b). In these (and subsequent) tables the ratios of 1 when $\phi = (2.29)$ and $L = 1$ reflect the identity $\hat{\theta}_A = \widetilde{\theta}$. Otherwise, though $\hat{\theta}_A$ is sometimes worse than $\widetilde{\theta}$ for small L, by $L = 4$ a clear, sometimes dramatic improvement was registered, especially in the MSE ratios.

The bimodal mixture normal is perhaps the qualitatively most different from the normal of all the distributions, and the potential for efficiency improvement greatest. Tables 7-18 confirm this. Nevertheless, except for $\lambda_0 = 0.8$ (in relative variance Tables 7, 11 and 15), $\hat{\theta}_A$ always beats OLS, to varying degrees. Some summary statistics based on all the Tables 3-18 are useful. Consider first the property of monotone improvement with increasing n or L (we do not count cases when, say, there is ultimate improvement without monotonicity). There is monotone improvement with increasing n in 84 (30 for $\widehat{\lambda}_A$, 54 for $\widehat{\beta}_A$) out of 160 places, with distribution (c) best and (a) worst. There is monotone improvement with increasing L in 104 (48 for $\widehat{\lambda}_A$, 56 for $\widehat{\beta}_A$) out of 196 places, with (b) best and (d) worst. In both instances, the number of such improvements was somewhat greater for $\lambda_0 = 0.4$ than $\lambda_0 = 0.8$. With respect to choice of ϕ, there is monotone improvement with increasing n in 23 of 64 places for (2.29) (omitting $L = 1$ of course) and 62 of 96 for (2.30).

(Tables 7-18 about here)

The disappointing aspects of Tables 7, 11 and 15 serve as a prelude to the results for $\hat{\theta}_B = (\widehat{\lambda}_B, \widehat{\beta}_B)^T$ when $\lambda_0 = 0.8$. What happens is that the second ("bias correction") component in s_L vastly overcompensates for the positive bias in $\widetilde{\lambda}$ seen in Table 1. The reason is apparent from (3.16). Overestimation of λ_0 not only increases the numerator but brings the denominator close to zero. In one place $\widehat{\lambda}_B$ beats OLS, and $\widehat{\beta}_A$ does so in 46, but these are out of 144 in each case, and overall the results are too poor to report. However, we present the results for $\lambda_0 = 0.4$, in Tables 19-26, combining relative variance and MSE in each table. Of most interest is comparison of $\hat{\theta}_B$ with $\hat{\theta}_A$. Of the 288 places, $\hat{\theta}_B$ does best in 124; 93 of these are relative variances, and 70 refer to $\widehat{\beta}_B$. The bias-correction is not very successful even when $\lambda_0 = 0.4$, with $\widetilde{\lambda}$ still largely to blame. There is monotone improvement with increasing n in 23 (11 for $\widehat{\lambda}_B$, 12 for $\widehat{\beta}_B$) out of 96 places, with distribution (c) best (again) and (e) worst, so $\hat{\theta}_B$ performs worse than $\hat{\theta}_A$ in this respect also. On the other hand, there is monotone improvement with increasing L in 56 (28 each for $\widehat{\lambda}_B$ and $\widehat{\beta}_B$) out of 92 places, with (b) best (again) and the others roughly equal. Again the choice (2.30) of ϕ fares better than (2.29) with respect to monotone improvement with increasing $n, 16$ to 7.

(Tables 19-26 about here)

Clearly $\hat{\theta}_D$, in particular a Newton approximation to the Gaussian MLE, will be similarly affected, relative to $\hat{\theta}_C$. Lee (2004), in his Monte Carlo, used a search algorithm to compute the Gaussian MLE itself, thereby not risking contamination by an initial estimate. However, the larger n, and especially k, the more expensive this approach becomes, and it could prove prohibitive, especially when W leads to a less tractable det $\{I_n - \lambda W\}$ than is the case with (1.4) (see Kelejian and Prucha (1999). Iteration from an initial estimate may then be preferable (which brings us back to $\hat{\theta}_D$). On the other hand, the present paper has stressed ahievement of asymptotic efficiency in a general setting, with a minimum of computation. In a given practical situation, this may not be the most relevant goal, and improvements might be desirable, perhaps especially to $\hat{\theta}_B$ and $\hat{\theta}_D$, by exercizing greater care in choice of $\widetilde{\theta}$ (possibly using one of the instrumental variables estimates in the literature), and continuing the iterations. This will incur greater computational expense, though updating of R does not arise. These and other issues might be examined in a subsequent, more thorough, Monte Carlo study. It is hoped that the present simulations have demonstrated that the computationally simple estimates $\hat{\theta}_A$ and $\hat{\theta}_B$, with their optimal asymptotic properties in a wide setting, offer sufficient promise to warrant such investigation and possible refinement, and empirical application.

APPENDIX: Proof of Theorem A

By the mean value theorem

$$\hat{\theta}_A - \theta_0 = \left(I_{k+1} - \frac{R^{-1}}{\tilde{\mathcal{I}}_L(\tilde{\theta}, \tilde{\sigma})}\bar{S}_{1L}\right)\left(\tilde{\theta} - \theta_0\right) - \frac{R^{-1}}{\tilde{\mathcal{I}}_L(\tilde{\theta}, \tilde{\sigma})}\left\{\bar{S}_{2L}(\tilde{\sigma} - \sigma_0) + r_L(\theta_0, \sigma_0)\right\} \quad \text{(A.1)}$$

where $\bar{S}_{1L}$ and $\bar{S}_{2L}$ are respectively obtained from $S_{1L}(\theta, \sigma) = (\partial/\partial\theta^T)r_L(\theta, \sigma)$ and $S_{2L}(\theta, \sigma) = (\partial/\partial\sigma)r_L(\theta, \sigma)$ after evaluating each row at some (possibly different) $\bar{\theta}$, $\bar{\sigma}$ such that $\|\bar{\theta} - \theta_0\| \le \left\|\tilde{\theta} - \theta_0\right\|$, $|\bar{\sigma} - \sigma_0| \le |\tilde{\sigma} - \sigma_0|$. Introduce the neighbourhood $\mathcal{N} = \left\{\theta, \sigma : \|\theta - \theta_0\| + \|\sigma - \sigma_0\| \le n^{-\frac{1}{2}}\right\}$. In view of Assumptions 2 and 3, the proof consists of showing that

$$\sup_{\mathcal{N}} \|S_{iL}(\theta,\sigma) - S_{iL}(\theta_0,\sigma_0)\| = o_p(n), \quad n = 1, 2, \tag{A.2}$$

$$\sup_{\mathcal{N}} \left|\tilde{\mathcal{I}}_L(\theta,\sigma) - \mathcal{I}_L(\theta_0,\sigma_0)\right| \to_p 0, \tag{A.3}$$

$$n^{-1}R \to_p \Omega, \tag{A.4}$$

$$\left[\tilde{\mathcal{I}}_L(\theta_0,\sigma_0)R\right]^{-1} S_{1L}(\theta_0,\sigma_0) \to_p I_{k+1}, \tag{A.5}$$

$$n^{-1}S_{L2}(\theta_0,\sigma_0) \to_p 0, \tag{A.6}$$

$$\tilde{\mathcal{I}}_L(\theta_0,\sigma_0) \to_p \mathcal{I}, \tag{A.7}$$

$$r_{1L} = o_p(n^{\frac{1}{2}}), \tag{A.8}$$

$$n^{-\frac{1}{2}}r_{2L} \to_d \mathcal{N}\left(0, \sigma_0\ \mathcal{I}\Omega\right), \tag{A.9}$$

where

$$r_{jL} = \sum_i \tilde{\psi}_{iL}(\theta_0,\sigma_0)E'_{ji}, \quad j = 1, 2, \tag{A.10}$$

in which $\sum_i$ denotes $\sum_{i=1}^n$,

$$(E'_{11}, ..., E'_{1n}) = E'_1 = -\sigma_0 \left(HG\varepsilon, 0\right)^T, \tag{A.11}$$

$$(E'_{21}, ..., E'_{2n}) = E'_2 = -\left(HGX\beta_0, HX\right)^T. \tag{A.12}$$

Notice that $r_L(\theta_0,\sigma_0) = r_{1L} + r_{2L}$, due to $E' = e' = E'_1 + E'_2$, since

$$\begin{aligned} e' &= -\left(G\left(l_n\mu_0 + X\beta_0 + \sigma_0\varepsilon\right), X\right)^T \\ &= -\left(\left(1-\lambda_0\right)^{-1} l_n\mu_0 + G\left(X\beta_0 + \sigma_0\varepsilon\right), X\right)^T. \end{aligned} \tag{A.13}$$

The proof of (A.9) is essentially as in Newey (1988, Theorem 2.3), Robinson (2005, Theorem 1) (the weaker conditions in the latter reference being reflected in Assumption 5). The only difference is the triangular array structure in the first element of r_{2L}. This makes no real difference to the proof that the $\tilde{\psi}_{iL}(\theta_0,\sigma_0)$ can be replaced by the $\psi(\varepsilon_i)$, whence $n^{-\frac{1}{2}}\sum_i \psi(\varepsilon_i)E'_{2i} \to_d N\left(0, \sigma_0^2\mathcal{I}\Omega\right)$ follows from a triangular-array central limit theorem (such as Lemma A.2 of Lee, 2002).

To prove (A.8), write

$$r_{1L} = \sigma_0 \left(a_1 + a_2 + a_3 + a_4, 0\right)^T \quad \text{(A.14)}$$
$$a_j = \sum_i b_{ji}\chi_{in}, \quad j = 1, ..., 4, \quad \text{(A.15)}$$
$$\chi_{in} = \varepsilon^T G^T \left(1_i - l_n/n\right), \quad \text{(A.16)}$$
$$b_{1i} = \psi(\varepsilon_i), \quad \text{(A.17)}$$
$$b_{2i} = \bar{\psi}^{(L)}\left(\varepsilon_i; a^{(L)}\right) - \psi(\varepsilon_i), \quad \text{(A.18)}$$
$$b_{3i} = \psi^{(L)}\left(\varepsilon_i; \hat{a}^{(L)}(\varepsilon)\right) - \bar{\psi}^{(L)}\left(\varepsilon_i; a^{(L)}\right), \quad \text{(A.19)}$$
$$b_{4i} = \tilde{\psi}_i^{(L)}\left(\theta_0, \sigma_0\right) - \psi^{(L)}\left(\varepsilon_i; \tilde{a}^{(L)}(\varepsilon)\right), \quad \text{(A.20)}$$

in which $\bar{\psi}^{(L)}\left(\varepsilon_i; a^{(L)}\right) = \bar{\phi}^{(L)}(\varepsilon)^T a^{(L)}$ (cf. (2.3)) and 1_i is the ith column of I_n
Define

$$t_{ijn} = 1_j^T G^T \left(1_i - l_n/n\right) = 1_j^T G^T 1_i - \sum_{\ell} 1_j^T G^T 1_\ell/n, \quad \text{(A.21)}$$

so that

$$\chi_{in} = \sum_j t_{ijn}. \quad \text{(A.22)}$$

Thus write

$$a_1 = \sum_i \psi\left(\varepsilon_i\right)\varepsilon_i t_{iin} + \sum_i \psi(\varepsilon_i) \sum_{j \neq i} \varepsilon_j t_{ijn}. \quad \text{(A.23)}$$

The absolute value of the first term has expectation bounded by

$$E\left|\psi(\varepsilon_i)\varepsilon_i\right| \left\{ \sum_i \left|1_i^T G^T 1_i\right| + \sum_i \sum_j \left|1_i^T G^T 1_j\right|/n \right\}. \quad \text{(A.24)}$$

For all i, j, Assumption 1 implies

$$1_i^T G^T 1_j = 1_j^T W S^{-1} 1_i = O(h_n^{-\frac{1}{2}}) \quad \text{(A.25)}$$

uniformly. (Lee (2002, p.258) gives (A.25) for $i = j$.) Since the first factor in (A.24) is bounded by $\left\{E\psi^2(\varepsilon_i)E\varepsilon_i^2\right\}^{\frac{1}{2}} < \infty$ it follows that (A.24) $= O_p(n/h_n)$. The second term in (A.23) has mean zero and variance $O(n/h_n)$. The proof of the latter statement is quickly obtained from that of Lemma A.1 of Lee (2002), which covers $\sum_i \sum_j \varepsilon_i \varepsilon_j t_{ijn}$: we can replace ε_i by $\psi(\varepsilon_i)$, noting

$$E\left(\varepsilon_i\psi(\varepsilon_i)\right) = -\int s f'(s) ds = E\left(\varepsilon_i^2\right), \quad \text{(A.26)}$$

(by integration-by-parts), Assumption 3, and we omit "diagonal terms" $i = j$ of Lee's (2002) statistic, to negligible effect, thus implying that we do not require

$E\left(\varepsilon_i^2\psi(\varepsilon_i)^2\right) < \infty$, which would correspond to his condition $E\left(\varepsilon_i^4\right) < \infty$. It follows that $a_1 = O_p\left(n/h_n + (n/h_n)^{\frac{1}{2}}\right) = o_p(n^{\frac{1}{2}})$.

Next write

$$a_2 = \sum_i \tau_i^{(L)} \varepsilon_i t_{iin} + \sum_i \tau_i^{(L)} \sum_{j \neq i} \varepsilon_j t_{ijn}, \tag{A.27}$$

where $\tau_i^{(L)} = \bar{\psi}^{(L)}\left(\varepsilon_i; a^{(L)}\right) - \psi(\varepsilon_i)$. The square of the first term has expectation bounded by

$$nE\left(\tau_i^{(L)2}\right)\sum_i t_{iin}^2, \tag{A.28}$$

using the Schwarz inequality. The expectation in (A.28) remains finite as $L \to \infty$, indeed it tends to zero (see Freud, 1971, pp.77-79), as is crucially used in the proof of (A.9) (see Newey (1988, p.329)). From (A.25), the summands in (A.28) are uniformly $O(h_n^{-2})$. From Assumption 3, the second term of (A.27) has mean zero and variance

$$\sum_i E\left(\tau_i^{(L)2}\right) E\left(\sum_{j\neq i} \varepsilon_j t_{jin}\right)^2 + \sum_i \sum_j E\left(\tau_i^{(L)}\varepsilon_i\right) E\left(\tau_j^{(L)}\varepsilon_j\right) t_{jin} t_{ijn}. \tag{A.29}$$

The first sum is $O\left(\sum_i \sum_j t_{jin}^2\right) = O\left(n^2/h_n^2\right)$, while the second is

$$O\left(\sum_i E\left(\tau_i^{(L)2}\right) E\varepsilon_i^2 \sum_j t_{jin}^2\right) = O\left(\sum_i \sum_j t_{jin}^2\right) = O\left(n^2/h_n^2\right) \tag{A.30}$$

also. We have shown that $E(a_2^2) = O(n^2/h_n^2)$, whence $a_2 = o_p(n^{\frac{1}{2}})$ from Assumption 1.

Since $\sum_i \chi_{in} \equiv 0$, we deduce that

$$a_3 = \left\{\hat{a}^{(L)}(\varepsilon) - a^{(L)}\right\}^T \sum_i \bar{\phi}^{(L)}(\varepsilon_i)\chi_{in}. \tag{A.31}$$

Proceeding as before, write

$$\sum_i \bar{\phi}^{(L)}(\varepsilon_i)\chi_{in} = \sum_i \bar{\phi}^{(L)}(\varepsilon_i)\varepsilon_i t_{iin} + \sum_{j\neq i} \bar{\phi}^{(L)}(\varepsilon_i) \sum_{j\neq i} \varepsilon_j t_{ijn}. \tag{A.32}$$

As in Robinson (2005), introduce the notation

$$\mu_c = 1 + E\left|\varepsilon_i\right|^c, \quad c \geq 0, \tag{A.33}$$

and

$$\rho_{uL} = CL, \qquad \text{if } u = 0, \tag{A.34}$$

$$= (CL)^{uL/\omega}, \quad \text{if } u > 0 \text{ and Assumption 5(ii) holds}, \tag{A.35}$$

$$= C^L, \qquad \text{if } u > 0 \text{ and Assumption 5(iii) holds}, \tag{A.36}$$

suppressing reference to C in ρ_{uL}. With $\|.\|$ denoting (Euclidean) norm, the squared norm of the first term on the right of (A.32) has expectation bounded by

$$\begin{aligned}\sum_i E\left\|\bar{\phi}^{(L)}(\varepsilon_i)\right\|^2\sum_j t_{jjn}^2 &\leq \frac{Cn^2}{h_n^2}\sum_{\ell=1}^{L}E\phi^{2\ell}(\varepsilon_i)\\ &\leq \frac{Cn^2}{h_n^2}\sum_{\ell=1}^{L}\mu_{2\kappa L}\\ &\leq \frac{Cn^2}{h_n^2}\rho_{2\kappa L},\end{aligned} \tag{A.37}$$

using Assumption 4, and then Lemma 9 of Robinson (2005). The second term of (A.32) has zero mean vector and covariance matrix

$$\begin{aligned}&\sum_i E\left\{\bar{\phi}^{(L)}(\varepsilon_i)\bar{\phi}^{(L)}(\varepsilon_i)^T\right\}\sum_{j\neq i}t_{ijn}^2\\ &+\sum_i\sum_j E\left\{\bar{\phi}^{(L)}(\varepsilon_i)\varepsilon_i\right\}E\left\{\bar{\phi}^{(L)}(\varepsilon_j)\varepsilon_j\right\}^T t_{ijn}t_{jin},\end{aligned} \tag{A.38}$$

which from earlier calculations has norm $O\left(n^2\rho_{2\kappa L}/h_n^2\right)$. Thus

$$\left\|\sum_i \bar{\phi}^{(L)}(\varepsilon_i)\chi_{in}\right\| = O_p\left(\frac{n\rho_{2\kappa L}^{\frac{1}{2}}}{h_n}\right). \tag{A.39}$$

This is dominated by the bound for the corresponding expression in Robinson (2005) - see the bottom of p.1820 and the bound top of p.1829, and note that there was an $n^{-\frac{1}{2}}$ factor incorporated. From the rest of the proof for A_{31} in the latter reference, it follows that $a_3 = o_p(n^{\frac{1}{2}})$.

Now write κ

$$\begin{aligned}a_4 \;=\; &\left\{\tilde{a}^{(L)}(E/\sigma_0)-\tilde{a}^{(L)}(\varepsilon)\right\}^T\sum_i \Phi^{(L)}(\varepsilon_i)\chi_{in}\\ &+\tilde{a}^{(L)}(E/\sigma_0)^T\sum_i\left\{\Phi^{(L)}(E_i/\sigma_0)-\Phi^{(L)}(\varepsilon_i)\right\}\chi_{in}.\end{aligned} \tag{A.40}$$

By the mean value theorem, with $\bar{\varepsilon} = n^{-1}\sum_i \varepsilon_i$,

$$\phi_\ell(E_i/\sigma_0)-\phi_\ell(\varepsilon_i) = -\bar{\varepsilon}\phi_\ell'(\varepsilon_i)+\frac{1}{2}\bar{\varepsilon}^2\phi_\ell''(\varepsilon_i^*)\bar{\varepsilon}^2, \tag{A.41}$$

where $|\varepsilon_i^*-\varepsilon_i| \leq |E_i/\sigma_0-\varepsilon_i| = |\bar{\varepsilon}|$. Now

$$\sum_i \phi_\ell'(\varepsilon_i)\chi_{in} = \sum_i\left\{\phi_\ell'(\varepsilon_i)-E\phi_\ell'(\varepsilon_i)\right\}\chi_{in}. \tag{A.42}$$

Proceeding much as before, and from Assumption 3 and (6.23) and Lemma 9 of Robinson (2005), this is $O_p\left(\left\{E\phi_\ell^2(\varepsilon_i)^2\right\}^{\frac{1}{2}}n/h_n\right) = O_p\left(\ell\mu_{2\kappa(\ell+K)}^{\frac{1}{2}}n/h_n\right)$.

Using $|\varepsilon_i^*| \leq |\varepsilon_i| + |\bar{\varepsilon}|$ and the c_r-inequality, and proceeding as in Robinson (2005, p. 1822),

$$\left|\sum_i \phi_\ell^{''}(\varepsilon_i^*)\chi_{in}\right| \leq C^{\kappa\ell+1}\ell^2 \sum_i \left\{1 + |\varepsilon_i|^{\kappa(\ell-1+2K)} + |\bar{\varepsilon}|^{\kappa(\ell-1+2K)}\right\} |\chi_{in}| . \quad \text{(A.43)}$$

The Scwarz inequality gives

$$E\,|\chi_{in}| \leq (E\chi_{in}^2)^{1/2} \leq (\sum_j t_{ijn}^2)^{1/2} \leq Cn^{1/2}/h_n, \quad \text{(A.44)}$$

$$E(|\varepsilon_i|^{\kappa(\ell-1+2K)}|\chi_{in}|) \leq C\mu_{2\kappa(\ell-1+2K)}^{1/2} n^{1/2}/h_n, \quad \text{(A.45)}$$

uniformly in i. With $\bar{\varepsilon} = O_p(n^{-1/2})$, we deduce from the above calculations and Lemma 9 of Robinson (2005) that

$$\left|\sum_i \left\{\Phi^{(L)}(E_i/\sigma_0) - \Phi^{(L)}(\varepsilon_i)\right\}\chi_{in}\right| = O_p(C^{2\kappa L+1}L^{1/2}\rho_{2\kappa L}^{1/2}n^{1/2}/h_n). \quad \text{(A.46)}$$

Comparison of this and (A.39) with the corresponding bounds in Robinson (2005, p.1823) indicates that again the latter dominate, so that the rest of the proof for A_{41} in the latter reference implies that $a_4 = o_p(n^{\frac{1}{2}})$ (indeed the remaining bounds needed are slightly better than those in Robinson (2005), because of the long range serial dependence complications there).

This completes the proof of (A.8), which is by far the most difficult and distinctive part of the Theorem proofs, due both to the simultaneity problem and the $n^{\frac{1}{2}}$ normalization. We thus omit the proof of (A.2)-(A.7), of which indeed (A.4) is in Lee (2002). □

Acknowledgements

I thank Afonso Goncalves da Silva for carrying out the simulations reported in Section 4. This research was supported by ESRC Grant R000239936.

Table 1
Monte Carlo Bias of OLS estimate of λ_0

λ_0	0.4			0.8		
n	96	198	392	96	198	392
(a)	0.0436	0.0995	0.1397	0.1373	0.1289	0.1376
(b)	0.0704	0.1029	0.1336	0.1399	0.1296	0.1362
(c)	0.0743	0.1125	0.1384	0.1410	0.1297	0.1364
(d)	0.0414	0.1102	0.1337	0.1370	0.1305	0.1365
(e)	0.0738	0.1056	0.1337	0.1411	0.1295	0.1365

Table 2
Monte Carlo Bias of OLS estimate of β_0

λ_0	0.4			0.8		
n	96	198	392	96	198	392
(a)	0.0103	-0.0114	-0.0125	0.0155	-0.0434	-0.0291
(b)	-0.0033	-0.0126	-0.0145	-0.0007	-0.0447	-0.0324
(c)	-0.0058	-0.0061	0.0011	-0.0029	-0.0381	-0.0163
(d)	0.0041	-0.0147	-0.0151	0.0095	-0.0462	-0.0321
(e)	0.0067	0.0036	-0.0074	0.0091	-0.0272	-0.0243

Table 3
Relative Variance, $\text{Var}(\widehat{\lambda}_A)/\text{Var}(OLS)$, Bimodal Mixture Normal (b)

	λ_0	0.4			0.8		
ϕ	$L \diagdown n$	96	198	392	96	198	392
	1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1	2	1.0116	0.9978	0.9712	1.0499	1.1077	1.0492
	4	0.3165	0.1395	0.1406	0.4688	0.5954	0.7249
	1	1.8846	2.3580	2.4790	2.1239	2.2482	2.3935
2	2	1.4788	2.0466	2.0889	2.9373	4.3919	4.5746
	4	0.2809	0.0894	0.0972	0.3306	0.4559	0.5405

Table 4
Relative MSE, $\text{MSE}(\widehat{\lambda}_A)/\text{MSE}(OLS)$, Bimodal Mixture Normal (b)

	λ_0	0.4			0.8		
ϕ	$L \diagdown n$	96	198	392	96	198	392
	1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1	2	1.0084	0.9958	0.9798	0.9737	0.9868	0.9884
	4	0.3080	0.1223	0.1051	0.1395	0.0859	0.0644
	1	1.9022	2.3733	2.4731	2.1503	2.3157	2.3996
2	2	1.4770	1.9917	2.0356	1.6565	1.8489	1.9748
	4	0.2732	0.0773	0.0702	0.0795	0.0480	0.0358

Table 5
Relative Variance, Var($\widehat{\beta}_A$)/Var(OLS), Bimodal Mixture Normal (b)

	λ_0	0.4			0.8		
ϕ	$L \diagdown n$	96	198	392	96	198	392
	1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1	2	1.0009	0.9986	0.9961	1.0058	1.0077	0.9967
	4	0.1822	0.1671	0.1453	0.2187	0.1965	0.1641
	1	2.2720	2.4539	2.5556	2.1362	2.3412	2.4658
2	2	1.7792	2.0118	2.2734	1.7425	2.0080	2.2346
	4	0.1163	0.1150	0.1109	0.1423	0.1303	0.1164

Table 6
Relative MSE, MSE($\widehat{\beta}_A$)/MSE(OLS), Bimodal Mixture Normal (b)

	λ_0	0.4			0.8		
ϕ	$L \diagdown n$	96	198	392	96	198	392
	1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1	2	1.0011	0.9992	0.9960	1.0059	1.0096	0.9962
	4	0.1823	0.1670	0.1444	0.2188	0.1934	0.1590
	1	2.2720	2.4527	2.5587	2.1362	2.3333	2.4710
2	2	1.7804	2.0153	2.2745	1.7434	2.0073	2.2307
	4	0.1163	0.1150	0.1102	0.1423	0.1282	0.1127

Table 7
Relative Variance, Var($\widehat{\lambda}_A$)/Var(OLS), Unimodal Mixture Normal (c)

	λ_0	0.4			0.8		
ϕ	$L \diagdown n$	96	198	392	96	198	392
	1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1	2	0.8428	0.8339	0.9142	1.6047	1.5192	1.4857
	4	0.5876	0.5565	0.5257	1.5906	1.3757	1.4045
	1	0.6517	0.5925	0.5392	1.2441	1.1066	1.0893
2	2	0.6763	0.5873	0.5417	1.4211	1.1593	1.1204
	4	0.6813	0.6088	0.5495	1.4868	1.2822	1.2687

Table 8
Relative MSE, MSE($\widehat{\lambda}_A$)/MSE(OLS), Unimodal Mixture Normal (c)

	λ_0	0.4			0.8		
ϕ	$L \diagdown n$	96	198	392	96	198	392
	1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1	2	0.8215	0.7971	0.8826	0.6733	0.7395	0.8501
	4	0.5736	0.5066	0.4587	0.5081	0.4072	0.3929
	1	0.6371	0.5586	0.4922	0.5690	0.4938	0.4491
2	2	0.6581	0.5511	0.4909	0.5653	0.4846	0.4454
	4	0.6639	0.5666	0.4909	0.5842	0.4877	0.4400

Table 9

Relative Variance, Var($\widehat{\beta}_A$)/Var(OLS), Unimodal Mixture Normal (c)

	λ_0	0.4			0.8		
ϕ	$L \diagdown n$	96	198	392	96	198	392
	1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1	2	0.7637	0.8297	0.9379	0.7717	0.8387	0.9377
	4	0.5952	0.5535	0.5717	0.6033	0.5675	0.5765
	1	0.6069	0.5739	0.5537	0.6233	0.5860	0.5596
2	2	0.6211	0.5777	0.5608	0.6339	0.5906	0.5659
	4	0.6413	0.5854	0.5659	0.6495	0.6078	0.5728

Table 10

Relative MSE, MSE($\widehat{\beta}_A$)/MSE(OLS), Unimodal Mixture Normal (c)

	λ_0	0.4			0.8		
ϕ	$L \diagdown n$	96	198	392	96	198	392
	1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1	2	0.7637	0.8295	0.9379	0.7717	0.8274	0.9364
	4	0.5952	0.5532	0.5719	0.6033	0.5603	0.5737
	1	0.6068	0.5736	0.5539	0.6232	0.5804	0.5575
2	2	0.6210	0.5773	0.5610	0.6339	0.5841	0.5637
	4	0.6420	0.5852	0.5661	0.6503	0.6001	0.5704

Table 11

Relative Variance, Var($\widehat{\lambda}_A$)/Var(OLS), Laplace (d)

	λ_0	0.4			0.8		
ϕ	$L \diagdown n$	96	198	392	96	198	392
	1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1	2	0.9508	0.9891	0.9861	1.0487	1.1435	1.0837
	4	0.8036	0.7566	0.7855	1.0686	1.3116	1.3987
	1	0.6927	0.6938	0.6990	0.8384	1.0125	1.0672
2	2	0.7034	0.7080	0.7039	0.8984	1.0898	1.1219
	4	0.7464	0.6360	0.6049	1.1042	1.4822	1.7349

Table 12

Relative MSE, MSE($\widehat{\lambda}_A$)/MSE(OLS), Laplace (d)

	λ_0	0.4			0.8		
ϕ	$L \diagdown n$	96	198	392	96	198	392
	1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1	2	0.9482	0.9800	0.9806	0.8952	0.9530	0.9711
	4	0.7980	0.7372	0.7404	0.6839	0.6907	0.6776
	1	0.6906	0.6787	0.6733	0.6631	0.6488	0.6357
2	2	0.6998	0.6870	0.6734	0.6412	0.6443	0.6318
	4	0.7405	0.5946	0.5395	0.5899	0.5150	0.4590

Table 13

Relative Variance, Var($\widehat{\beta}_A$)/Var(OLS), Laplace (d)

	λ_0	0.4			0.8		
ϕ	$L \diagdown n$	96	198	392	96	198	392
	1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1	2	0.9771	0.9653	0.9823	0.9843	0.9661	0.9828
	4	0.8441	0.7839	0.7763	0.8609	0.7968	0.7861
	1	0.7319	0.6929	0.6781	0.7439	0.7110	0.6909
2	2	0.7397	0.6955	0.6781	0.7583	0.7130	0.6914
	4	0.7470	0.6471	0.6153	0.7864	0.6957	0.6436

Table 14

Relative MSE, MSE($\widehat{\beta}_A$)/MSE(OLS), Laplace (d)

	λ_0	0.4			0.8		
ϕ	$L \diagdown n$	96	198	392	96	198	392
	1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1	2	0.9773	0.9640	0.9826	0.9849	0.9609	0.9829
	4	0.8446	0.7830	0.7765	0.8618	0.7904	0.7839
	1	0.7322	0.6928	0.6788	0.7446	0.7092	0.6909
2	2	0.7400	0.6947	0.6794	0.7590	0.7087	0.6924
	4	0.7477	0.6461	0.6171	0.7875	0.6867	0.6426

Table 15

Relative Variance, Var($\widehat{\lambda}_A$)/Var(OLS), Student t_5 (e)

	λ_0	0.4			0.8		
ϕ	$L \diagdown n$	96	198	392	96	198	392
	1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1	2	0.9598	0.8858	0.9760	1.3243	0.9844	1.1553
	4	0.9225	0.7550	0.8714	1.5289	1.1056	1.4613
	1	0.7993	0.7593	0.8510	1.2130	1.0426	1.3446
2	2	0.8270	0.7716	0.8487	1.4044	1.0608	1.3821
	4	0.9127	0.7722	0.8678	1.7733	1.1332	1.5380

Table 16

Relative MSE, MSE($\widehat{\lambda}_A$)/MSE(OLS), Student t_5 (e)

	λ_0	0.4			0.8		
ϕ	$L \diagdown n$	96	198	392	96	198	392
	1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1	2	0.9497	0.8930	0.9674	0.8999	0.9239	0.9471
	4	0.9084	0.7553	0.8017	0.8251	0.7681	0.7261
	1	0.7910	0.7555	0.7848	0.7713	0.7502	0.7123
2	2	0.8156	0.7645	0.7834	0.7780	0.7486	0.7117
	4	0.9025	0.7690	0.7942	0.8355	0.7675	0.7154

Table 17

Relative Variance, Var($\widehat{\beta}_A$)/Var(OLS), Student t_5 (e)

ϕ	λ_0 / $L \diagdown n$	0.4: 96	0.4: 198	0.4: 392	0.8: 96	0.8: 198	0.8: 392
	1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1	2	0.9188	0.9485	0.9690	0.9245	0.9508	0.9694
	4	0.8735	0.8558	0.8035	0.8786	0.8646	0.8098
	1	0.8240	0.8178	0.7820	0.8285	0.8240	0.7930
2	2	0.8385	0.8318	0.7841	0.8507	0.8403	0.7942
	4	0.9431	0.8807	0.7882	0.9446	0.8862	0.8002

Table 18

Relative MSE, MSE($\widehat{\beta}_A$)/MSE(OLS), Student t_5 (e)

ϕ	λ_0 / $L \diagdown n$	0.4: 96	0.4: 198	0.4: 392	0.8: 96	0.8: 198	0.8: 392
	1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1	2	0.9187	0.9485	0.9694	0.9244	0.9504	0.9703
	4	0.8733	0.8557	0.8045	0.8785	0.8630	0.8112
	1	0.8239	0.8178	0.7830	0.8284	0.8226	0.7944
2	2	0.8383	0.8318	0.7853	0.8506	0.8384	0.7957
	4	0.9428	0.8811	0.7891	0.9441	0.8825	0.8000

Table 19

Relative Variance and MSE of $\widehat{\lambda}_B$, Bimodal Mixture Normal (b), $\lambda_0 = 0.4$

ϕ	$L \diagdown n$	Var: 96	Var: 198	Var: 392	MSE: 96	MSE: 198	MSE: 392
	1	0.0697	0.1392	0.1263	1.0148	1.7253	2.1979
1	2	0.0859	0.1615	0.1363	0.9207	1.6213	2.1001
	4	0.0878	0.1091	0.1179	0.1195	0.1375	0.1265
	1	0.8665	1.2386	1.1801	6.5737	12.9658	18.5581
2	2	1.0812	1.7573	1.8766	3.9639	8.9420	13.9850
	4	0.1001	0.0810	0.0911	0.1225	0.0986	0.0903

Table 20

Relative Variance and MSE of $\widehat{\beta}_B$, Bimodal Mixture Normal (b), $\lambda_0 = 0.4$

ϕ	$L \diagdown n$	Var: 96	Var: 198	Var: 392	MSE: 96	MSE: 198	MSE: 392
	1	0.9205	1.0670	1.0146	0.9322	1.0723	1.0082
1	2	0.9244	1.0629	1.0105	0.9368	1.0664	1.0041
	4	0.1752	0.1693	0.1447	0.1761	0.1689	0.1443
	1	2.0524	2.8491	2.7023	2.1118	2.9154	2.6935
2	2	1.6370	2.3079	2.3982	1.6769	2.3323	2.3881
	4	0.1141	0.1161	0.1106	0.1145	0.1158	0.1100

Table 21
Relative Variance and MSE of $\widehat{\lambda}_B$, Unimodal Mixture Normal (c), $\lambda_0 = 0.4$

ϕ	$L \diagdown n$	Var 96	Var 198	Var 392	MSE 96	MSE 198	MSE 392
1	1	0.7767	0.4680	0.4227	2.7241	2.3638	2.6039
	2	0.8157	0.5126	0.4754	2.1939	1.8923	2.2458
	4	0.7426	0.5508	0.5250	1.6199	1.2681	1.2840
2	1	1.1995	0.6172	0.5276	2.3637	1.3202	1.1512
	2	1.1122	0.6027	0.5278	2.2021	1.2739	1.1342
	4	0.7258	0.5139	0.4919	1.4882	1.0777	1.0517

Table 22
Relative Variance and MSE of $\widehat{\beta}_B$, Unimodal Mixture Normal (c), $\lambda_0 = 0.4$

ϕ	$L \diagdown n$	Var 96	Var 198	Var 392	MSE 96	MSE 198	MSE 392
1	1	0.9455	1.0254	1.0339	0.9653	1.0421	1.0435
	2	0.7365	0.8505	0.9685	0.7498	0.8725	0.9771
	4	0.5785	0.5561	0.5835	0.5873	0.5653	0.5888
2	1	0.5926	0.5783	0.5687	0.6028	0.5884	0.5737
	2	0.6057	0.5824	0.5760	0.6153	0.5925	0.5810
	4	0.6272	0.5861	0.5796	0.6386	0.5960	0.5845

Table 23
Relative Variance and MSE of $\widehat{\lambda}_B$, Laplace (d), $\lambda_0 = 0.4$

ϕ	$L \diagdown n$	Var 96	Var 198	Var 392	MSE 96	MSE 198	MSE 392
1	1	0.1770	0.2231	0.2026	1.2049	1.7983	2.2863
	2	0.1863	0.2444	0.2184	1.0836	1.6848	2.1825
	4	0.2149	0.2834	0.2738	0.8740	1.2912	1.6059
2	1	0.2879	0.2528	0.2259	0.8644	1.0467	1.1708
	2	0.2691	0.2556	0.2287	0.8038	1.0129	1.1459
	4	0.2093	0.3239	0.3242	0.6072	0.9021	1.0168

Table 24
Relative Variance and MSE of $\widehat{\beta}_B$, Laplace (d), $\lambda_0 = 0.4$

ϕ	$L \diagdown n$	Var 96	Var 198	Var 392	MSE 96	MSE 198	MSE 392
1	1	0.9210	1.0588	1.0240	0.9297	1.0609	1.0168
	2	0.9092	1.0219	1.0043	0.9151	1.0255	0.9972
	4	0.7942	0.8188	0.7880	0.7973	0.8203	0.7824
2	1	0.6809	0.7219	0.6865	0.6842	0.7219	0.6817
	2	0.6929	0.7224	0.6876	0.6956	0.7231	0.6829
	4	0.7114	0.6687	0.6191	0.7124	0.6688	0.6152

Table 25

Relative Variance and MSE of $\widehat{\lambda}_B$, t_5 (e), $\lambda_0 = 0.4$

		Var			MSE		
ϕ	$L \diagdown n$	96	198	392	96	198	392
	1	0.2615	0.2276	0.2236	1.9481	1.8278	2.3424
1	2	0.2874	0.2411	0.2736	1.7639	1.6659	2.1911
	4	0.2938	0.2795	0.3410	1.4688	1.3963	1.8408
	1	0.4448	0.3012	0.3692	1.7832	1.3997	1.7493
2	2	0.4248	0.3042	0.3742	1.6570	1.3534	1.7110
	4	0.3506	0.3015	0.3655	1.2127	1.1610	1.5755

Table 26

Relative Variance and MSE of $\widehat{\beta}_B$, t_5 (e), $\lambda_0 = 0.4$

		Var			MSE		
ϕ	$L \diagdown n$	96	198	392	96	198	392
	1	0.9216	1.0587	1.0305	0.9300	1.0817	1.0305
1	2	0.8496	0.9941	0.9974	0.8572	1.0156	0.9969
	4	0.8151	0.8931	0.8237	0.8221	0.9098	0.8228
	1	0.7657	0.8614	0.7983	0.7726	0.8782	0.7974
2	2	0.7822	0.8730	0.8019	0.7894	0.8890	0.8009
	4	0.8860	0.9198	0.8068	0.8922	0.9371	0.8058

References

Arbia, G., 2006 Spatial econometrics: statistical foundations and applications to regional analysis. (Springer-Verlag, Berlin).

Beran, R., 1976 Adaptive estimates for autoregressive processes. Annals of the Institute of Statistical Mathematics 26, 77-89.

Bickel, P., 1982 On adaptive estimation. Annals of Statistics 10, 647-671.

Case, A.C., 1991 Neighbourhood influence and technological change. Regional Science and Urban Economics 22, 491-508.

Freud, G., 1971 Orthogonal polynomials. (Pergamon Press, Oxford).

Kelejian, H.H., Prucha, I.R., 1998 A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. Journal of Real Estate Finance and Economics 17, 99-121.

Kelejian, H.H., Prucha, I.R., 1999 A generalized moments estimator for the autoregressive parameter in a spatial model. International Economic Review 40, 509-533.

Kelejian, H.H., Prucha, I.R., Yuzefovich, Y., 2003 Instrumental variable estimation of a spatial autoregressive model with autoregressive disturbances: large and small sample results. Preprint.

Lee, L.F., 2002 Consistency and efficiency of least squares estimation for mixed regressive, spatial autoregressive models. Econometric Theory 18, 252-277.

Lee, L.F., 2003 Best spatial two-stage least squares estimators for a spatial autoregressive model with autoregressive disturbances. Econometric Reviews 22, 307-335.

Lee, L.F., 2004 Asymptotic distributions of quasi-maximum likelihood estimates for spatial autoregressive models. Econometrica 72, 1899-1925.

Newey, W.K., 1988 Adaptive estimation of regression models via moment restrictions. Journal of Econometrics 38, 301-339.

Robinson, P.M., 1988 The stochastic difference between econometric statistics. Econometrica 56, 531-548.

Robinson, P.M., 2005 Efficiency improvements in inference on stationary and nonstationary fractional time series. Annals of Statistics 33, 1800-1842.

Stone, C.J., 1975 Adaptive maximum likelihood estimators of a location parameter. Annals of Statistics 2, 267-289.

Whittle, P., 1954 On stationary processes in the plane. Biometrika 41, 434-449.