A Maximum Likelihood Model for Econometric Estimation with Spatial Series

L.W.HEPPLE
University of Bristol

1 The problem of dependence

The standard linear regression model is

$$y = \mathbf{X}\beta + u \ , \tag{1}$$

where

y is the $n \times 1$ vector of observations on the dependent variable;
$\mathbf{X}$ is the $n \times k$ matrix of observations on the k independent variables;
β is the $k \times 1$ vector of regression coefficients; and
u is the $n \times 1$ vector of errors.

We assume that the error terms, u, are independent of one another, so that the variance–covariance matrix of errors is diagonal (and is also assumed to be homoscedastic):

$$\mathrm{E}(uu') = \sigma^2\mathrm{I} \ , \tag{2}$$

where I is the unit matrix. If the errors are in fact dependent on each other in some way, so that

$$\mathrm{E}(uu') = \sigma^2\mathbf{V} \ , \tag{3}$$

where $\mathbf{V}$ is nondiagonal, then the least-squares estimates of β may be shown to be inefficient and the standard errors and significance tests biased, except in certain very specialised instances. In estimation and model building with time-series data, econometricians generally recognise that the formulation in equation (3) is more plausible than that in equation (2), and this is frequently confirmed by tests for serial correlation in the errors. It is also increasingly common for econometricians not simply to test for serial correlation, but to incorporate the dependence within u into the estimation procedure. If V is known *a priori* then the generalised least-squares estimator:

$$\hat{\beta} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}y \ , \tag{4}$$

provides efficient estimates and facilitates valid hypothesis-testing (Theil, 1971). Only rarely is V completely known. If, however, V depends on only a few unknown parameters, these may be estimated from the sample data, together with $\hat{\beta}$. A commonly hypothesised generating process for the dependence in u is the first-order temporal autoregressive model:

$$u_t = \rho u_{t-1} + e_t \ , \qquad t = 1, 2, ..., n \ , \tag{5}$$

where e is a random sample from an independent normal distribution with mean zero and variance σ^2, which we denote by $e \sim N(0, \sigma^2\mathbf{I})$. For quarterly economic data, a fourth-order process is perhaps more plausible (Wallis, 1971). An alternative specification is a moving-average process, and complex mixed moving-average autoregressive processes may also be considered (Box and Jenkins, 1970; Dhrymes, 1971). The first-order autoregressive model is, however, still the formulation most frequently used in applied econometric work.

Equation (5) may be rewritten in matrix terms as

$$u = \rho \mathbf{M} u + e \tag{6}$$

where $\mathbf{M}$ is the lag-operator matrix, that is it contains zeros apart from ones on the first subdiagonal. Denoting $\mathbf{P} = \mathbf{I} - \rho\mathbf{M}$ in equation (6), except for the very first element p_{11}, which is $(1-\rho^2)^{1/2}$, and is introduced to incorporate the variance of the first observation (see Dhrymes, 1971; or Hildreth, 1969), we may show that $\mathbf{P}'\mathbf{P} = \mathbf{V}^{-1}$. For the temporal autoregressive model, it may then be shown that minimising $e'e$, which may also be written $u'\mathbf{V}^{-1}u$, produces consistent and hence asymptotically unbiased and efficient estimates (Hildreth and Lu, 1960). This minimising may be achieved by direct nonlinear least-squares estimation of $\hat{\beta}$ and $\hat{\rho}$ simultaneously.

Alternatively a number of multistage linear least-squares estimation methods have been derived. Two well-known procedures of this type, due to Durbin (1960), and Cochrane and Orcutt (1949) are described in Johnston (1972) and most econometrics texts. This group of estimators that minimise $u'\mathbf{V}^{-1}u$ are sometimes known as minimum chi-square estimators (Dhrymes, 1971). Similar methods have also been devised for incorporating temporal dependence into simultaneous equation systems (Goldfeld and Quandt, 1972).

The importance of incorporating autocorrelated errors into the estimation procedure has been neatly demonstrated in econometric studies of United Kingdom inflationary problems by Wallis (1971) and Godfrey (1971), where the inclusion of autocorrelation in both single-equation and simultaneous-equation models overturns the ordinary least-squares inferences on the effectiveness of incomes policy.

2 Estimation with spatial series

In regional science, urban economics, economic geography, and econometrics generally, large numbers of regression models of the form of equation (1) are estimated on the basis of a spatial cross-section of n regions, states, or counties. Increasingly, simultaneous-equation techniques are also being employed with spatial series (for example Lee, 1970; Steinnes and Fisher, 1974). The presence of *spatial* autocorrelation or dependence in both variables and regression residuals is increasingly recognised and statistically demonstrated (Cliff and Ord, 1973; Hepple, 1974a; 1974b), and it is no more plausible for spatial series than for time series that the error variance-covariance matrix of a regression model will be diagonal. Until recently, however, the problem was almost totally neglected [see Hepple (1974b) for a history of this neglect and recent research]. Tests for spatial autocorrelation amongst regression residuals are now available (Cliff and Ord, 1973; Hepple, 1974a), and the severe effects of neglecting autocorrelated errors in spatial regression have been studied (Hepple, 1974a); it is important to develop estimation procedures that incorporate *spatial* dependence into econometric estimation of both single-equation and simultaneous-equation models.

As in the temporal case, it is necessary to decompose **V** so that it depends on only a few parameters. The commonest form of spatial autocorrelation is positive dependence between adjacent regions, so that a generating process incorporating dependence on immediate neighbours should give a valid and parsimonious representation of **V**. One such representation is a first-order spatial autoregressive model of the form of equation (6), where **M** is now some form of spatial lag-operator. Let **W** be the $n \times n$ matrix of contiguities between regions, with entry $w_{ij} = 1$ if regions i and j are contiguous, and $w_{ij} = 0$ otherwise, and let **M** be matrix **W** with the rows each scaled to sum to unity. This gives a generating process in which each region is a linear function of the average contiguous value. **M** can also, of course, incorporate prior weightings based on other information about interregional linkages and flows. Other specifications of the generating process, such as spatial moving averages and Box-Jenkins models, are also possible, but as in the time-series literature, the autoregressive model is a useful starting point.

Estimation of the regression model with autocorrelated errors is more difficult in the spatial case than with temporal data. The least-squares or minimum chi-square procedures for time series do not have valid analogues in the spatial-series case. The validity of least-squares or minimum chi-square estimators in time-series models springs from the asymptotic equivalence of these procedures with maximum-likelihood estimators. In spatial models this equivalence does not exist, and minimum chi-square estimators are not consistent. The reasons for this will be examined in subsequent sections.

Since least-squares procedures are invalid, the more general context of maximum-likelihood estimators has to be investigated. Coefficient estimates that maximise the likelihood function have desirable properties of consistency (and hence asymptotic unbiasedness and efficiency) and asymptotic multivariate normality. These properties are set out in Cramer (1945) and Goldfeld and Quandt (1972). Under suitable regularity conditions, such as the existence of the relevant derivatives, the estimates may be obtained by finding values of the coefficients that fulfil the second-order conditions for a maximum when first derivatives are set to zero. The set of values giving rise to the largest maximum constitutes the maximum-likelihood estimates. The variance–covariance matrix of the coefficients, **R**, may then be estimated by the negative inverse of the probability limit of the Hessian matrix (the matrix of second-order partial derivatives). **R** may then be used, together with the property of asymptotic multivariate normality, to provide asymptotic significance tests of the coefficients. In the following two sections the likelihood function, derivatives, and asymptotic distribution theory are developed for the regression model with spatially autocorrelated errors.

3 The likelihood function

The density function of a standard normal independent variable, e, is:

$$(2\pi)^{-n/2}\exp(-\tfrac{1}{2}e'e)\ , \qquad e \sim N(0, \mathrm{I}_n)\ . \tag{7}$$

Transforming to another variable, u, where $u = B+\mathrm{C}e$, gives:

$$(2\pi)^{-n/2}|\mathrm{C}^{-1}|\exp\{-\tfrac{1}{2}(u-B)'(\mathrm{C}^{-1})'\mathrm{C}^{-1}(u-B)\}\ . \tag{8}$$

The term $|\mathrm{C}^{-1}|$, the Jacobian of the transformation, is the determinant of the matrix of partial derivatives of the transformation, with typical element $\partial e_j/\partial u_i$. In a transformation such as $y = \mathrm{X}\beta+e$, C^{-1} is simply an identity matrix, so the term disappears. For $y = \sigma e$, where σ is scalar, then C^{-1} is diagonal. Where σ_u^2 is not unity, so that σ is the standard deviation, the density becomes:

$$(2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}u'u\right)\ . \tag{9}$$

For an observed series, this gives the likelihood function. In most models $|\mathrm{C}^{-1}|$ is either unity or a simple diagonal form, as with the variance transformation. Only if u_j is related to elements of e other than e_j does the Jacobian become a problem.

It is usual for convenience to maximise the log-likelihood function. For formulation (9) the log-likelihood function is

$$L = -\frac{n}{2}\ln 2\pi-\frac{n}{2}\ln\sigma^2-\frac{1}{2\sigma^2}u'u\ , \tag{10}$$

and for the normal multiple-regression model with independent errors:

$$L = -\frac{n}{2}\ln 2\pi-\frac{n}{2}\ln\sigma^2-\frac{1}{2\sigma^2}(y-\mathrm{X}\beta)'(y-\mathrm{X}\beta)\ . \tag{11}$$

Minimisation of the quadratic component will maximise the log-likelihood. This minimisation is the equivalent of least-squares minimisation.

Now consider the likelihood function for models with autocorrelated errors, where the Jacobian element may be important. For both the spatial and temporal autoregressive formulations, it follows from equation (6) that $\boldsymbol{u} = \mathbf{P}^{-1}\boldsymbol{e}$. Hence in both these models $\mathbf{C}^{-1} = \mathbf{P}$, and if we recall that $\mathbf{P}'\mathbf{P} = \mathbf{V}^{-1}$, the log-likelihood function is then:

$$L = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 + \ln|\mathbf{P}| - \frac{1}{2\sigma^2}[(y-\mathbf{X}\beta)'\mathbf{V}^{-1}(y-\mathbf{X}\beta)] . \tag{12}$$

In the time-series case, $\mathbf{P}$ is a matrix with diagonal entries of unity [except for p_{11} which is $(1-\rho^2)^{1/2}$], and $-\rho$ on the first subdiagonal, with all the other elements zero. $\mathbf{P}$ is therefore a triangular matrix. The determinant of a triangular matrix is the product of its diagonal elements. Hence, if we ignore p_{11}, $|\mathbf{C}^{-1}|$ is unity, if we assume that σ^2 has already been separated out, so that least-squares estimation (minimisation of the quadratic component) is equivalent to maximum-likelihood estimation. If we recognise that the first element of matrix $\mathbf{P}$ is not unity but $(1-\rho^2)^{1/2}$, then the determinant $|\mathbf{C}^{-1}|$ is $(1-\rho^2)^{1/2}$. However, as sample size n increases, this term does not alter, its magnitude becoming swamped by the quadratic component of the likelihood function, so that asymptotically it is negligible (this is further clarified in section 6).

In the spatial series case the same results do not apply. In the spatial model $\mathbf{C}^{-1}$ will not be a triangular matrix: region j influences region i, which in turn influences region j. Both m_{ij} and m_{ji}, elements of $\mathbf{M}$, will be nonzero if regions i and j are adjacent, but they will probably not be equal. The Jacobian is neither negligible nor invariant with sample size, and does not collapse to a simple quantity. Whittle (1954), writing in a spectral-methods context, was the first to note the problem of the non-vanishing Jacobian. It was also noted in a biometric context by Mead (1967). Independently of the present work, Cliff and Ord (1973) have presented the likelihood in a regression context and provided an empirical illustration. There has, however, been no study of the structure and properties of the maximum-likelihood estimation, the role of the Jacobian, distribution theory, or computational techniques. These are developed here.

4 Derivatives and asymptotic distribution

The first partial derivatives of the log-likelihood function in equation (12) for the spatial model are:

$$\frac{\partial L}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\boldsymbol{u}'\mathbf{V}^{-1}\boldsymbol{u} , \tag{13}$$

$$\frac{\partial L}{\partial\beta} = \frac{1}{\sigma^2}\mathbf{X}'\mathbf{V}^{-1}\boldsymbol{u} , \tag{14}$$

$$\frac{\partial L}{\partial\rho} = -\mathrm{tr}\mathbf{P}^{-1}\mathbf{M} + \frac{1}{2\sigma^2}(\mathbf{P}'\mathbf{M}+\mathbf{M}'\mathbf{P}) . \tag{15}$$

The probability limits of these derivatives may be shown to be zero (Hepple, 1974a; in preparation). The second partial derivatives are:

$$\frac{\partial^2 L}{\partial\sigma^2\partial\sigma^2} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\boldsymbol{u}'\mathbf{V}^{-1}\boldsymbol{u} \,, \tag{16}$$

$$\frac{\partial^2 L}{\partial\beta\partial\beta} = -\frac{1}{\sigma^2}\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} \,, \tag{17}$$

$$\frac{\partial^2 L}{\partial\rho\partial\rho} = -\mathrm{tr}(\mathbf{P}^{-1}\mathbf{M})^2 - \frac{1}{\sigma^2}\boldsymbol{u}'\mathbf{M}'\mathbf{M}\boldsymbol{u} \,, \tag{18}$$

$$\frac{\partial^2 L}{\partial\sigma^2\partial\beta} = -\frac{1}{\sigma^4}\mathbf{X}'\mathbf{V}^{-1}\boldsymbol{u} \,, \tag{19}$$

$$\frac{\partial^2 L}{\partial\sigma^2\partial\rho} = -\frac{1}{2\sigma^4}\boldsymbol{u}'(\mathbf{P}'\mathbf{M}+\mathbf{M}'\mathbf{P})\boldsymbol{u} \,, \tag{20}$$

$$\frac{\partial^2 L}{\partial\beta\partial\rho} = -\frac{1}{\sigma^2}\mathbf{X}'(\mathbf{P}'\mathbf{M}+\mathbf{M}'\mathbf{P})\boldsymbol{u} \,. \tag{21}$$

The probability limits (plim) of these derivatives are:

$$\operatorname*{plim}_{n\to\infty}\frac{1}{n}\frac{\partial^2 L}{\partial\sigma^2\partial\sigma^2} = -\frac{1}{2\sigma^4} \,, \tag{22}$$

$$\operatorname*{plim}_{n\to\infty}\frac{1}{n}\frac{\partial^2 L}{\partial\beta\partial\beta} = -\frac{1}{n\sigma^2}\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} \,, \tag{23}$$

$$\operatorname*{plim}_{n\to\infty}\frac{1}{n}\frac{\partial^2 L}{\partial\rho\partial\rho} = -\frac{1}{n}\mathrm{tr}\,\mathbf{A}^2 - \frac{1}{n}\mathrm{tr}\,\mathbf{A}'\mathbf{A} \,, \tag{24}$$

where $\mathbf{A} = \mathbf{MP}^{-1}$, with

$$\operatorname*{plim}_{n\to\infty}\frac{1}{n}\frac{\partial^2 L}{\partial\sigma^2\partial\rho} = -\frac{\mathrm{tr}\,\mathbf{A}}{n\sigma^2} \,, \tag{25}$$

and the probability limits of the cross-partials between β and σ^2, and between β and ρ, are zero. Let us denote the matrix of these probability limits by Ω.

$$\operatorname*{plim}_{n\to\infty}\frac{1}{n}\frac{\partial^2 L}{\partial\theta'\partial\theta} = \Omega \tag{26}$$

where $\theta = (\sigma^2, \beta, \rho)$.

The negative inverse of this matrix Ω is the asymptotic variance–covariance matrix of the maximum likelihood estimates. Ω may be partitioned so that $\hat{\beta}$ may be tested separately from $\hat{\rho}$ and σ^2. These results may be compared with those for temporal models (see for example Dhrymes, 1971; Hildreth, 1969). It is the elements containing ρ that differ from temporal models. The second partial for ρ in equation (24) naturally differs, and the probability limit for the cross-partial, equation (25), is zero in the temporal model.

5 Computation of maximum-likelihood estimates

The likelihood function is highly nonlinear, and numerical optimisation techniques are needed to determine the maximum. A wide variety of such techniques are now available (Fletcher, 1970; Himmelblau, 1972), some of which, such as Newton-type methods, require repeated evaluation of both first and second partial derivatives, others only the first partial derivatives (for example steepest ascent and Davidon's method), and some require neither set of derivatives, as in Powell's conjugate gradient algorithm (Powell, 1965). Where explicit expressions for the derivatives are available it is normally most efficient to employ them in the maximisation. However, if computation of these derivatives, even if available, is complex and time consuming—as in the present case where inversion of $n \times n$ matrices would be involved—methods such as Powell's may be faster and less subject to round-off error.

Powell's conjugate gradient algorithm is based on solving a sequence of one-dimensional maximisation problems. If there are k unknown coefficients to be estimated, each iteration basically consists of maximising the function sequentially in k conjugate directions. After each iteration either the k directions are retained, or one is abandoned and a new conjugate direction calculated. So for a quadratic function a maximum of k iterations are needed to find the maximum. The algorithm also has good convergence properties for more general functions and has found extensive use in econometric research (for example Hendry, 1971; Hendry and Trevedi, 1970). It is the method used in the empirical illustration below.

The major difficulty in numerical maximisation of the likelihood function is the necessity of evaluating the $n \times n$ determinant $|\mathbf{P}|$ at each step. Ostensibly this severely limits the general availability of the method. However, a simplification is possible for the model with first-order spatial autoregressive errors. (Simplifications are also possible for a number of other generating processes.) Recalling that $\mathbf{P} = \mathbf{I} - \rho\mathbf{M}$, we denote the eigenvalues of $\mathbf{P}$ by λ_i, $i = 1, ..., n$, and the eigenvalues of $\mathbf{M}$ by μ_i, $i = 1, ..., n$. It may be proved (see Hepple, 1974a; in preparation) that for each nonzero eigenvalue of $\mathbf{M}$, μ_i, there is a corresponding eigenvalue of $\mathbf{P}$, λ_i, equal to $1 - \rho\mu_i$. Since the determinant of a matrix is the product of its eigenvalues:

$$|\mathbf{P}| = \prod_{i=1}^{n} \lambda_i = \prod_{i=1}^{n} (1 - \rho\mu_i) \,. \tag{27}$$

The advantage of this formulation is that the eigenvalues of $\mathbf{M}$ may be determined once and for all at the beginning of the estimation. Unless the contiguity matrix $\mathbf{W}$ is 'regular', with each region in the lattice having the same number of adjacencies, $\mathbf{M}$ will not be a symmetric matrix. Hence the eigenvalues of a general real matrix must be extracted. This may be done by reducing the matrix to an upper-Hessenberg form and then using Francis' QR algorithm to extract the eigenvalues (Wilkinson, 1965). This procedure is quite practicable even for large matrices: for the American illustration below, the eigenvalues of a 49 x 49 nonsymmetric matrix were extracted. Given these eigenvalues of **M**, $|\mathbf{P}|$ is easily obtained at each function evaluation.

6 The role of the Jacobian

It is useful to explore further the role of the Jacobian term, $|\mathbf{P}|$, in the estimation. Concentrating the log-likelihood function presented in equation (12) by eliminating σ^2 gives

$$L^* = -\frac{n}{2}(\ln 2\pi + 1) + \ln|\mathbf{P}| - \frac{n}{2}\ln \boldsymbol{u}'\mathbf{V}^{-1}\boldsymbol{u} , \tag{28}$$

$$= -\frac{n}{2}(\ln 2\pi + 1) - \frac{n}{2}\ln\frac{\boldsymbol{u}'\mathbf{V}^{-1}\boldsymbol{u}}{|\mathbf{P}|^{2/n}} . \tag{29}$$

Maximising the log-likelihood function is thus equivalent to globally minimising $\boldsymbol{u}'\mathbf{V}^{-1}\boldsymbol{u}/(|\mathbf{P}|^{2/n})$. This formulation is directly equivalent to expression (39) in Whittle's paper on a spectral approach to spatial estimation, where in his notation KU is the expression to be minimised (Whittle, 1954, p.44). K is $|\mathbf{P}|^{-2/n}$ and U is $\boldsymbol{u}'\mathbf{V}^{-1}\boldsymbol{u}$. Using this notation, K acts essentially as a penalty or weighting function on the sum of squares component $\boldsymbol{u}'\mathbf{V}^{-1}\boldsymbol{u}$. Least-squares estimates simply set $K = 1$ for all values of ρ.

Equation (29) also allows further clarification of the equivalence of maximum-likelihood and minimum chi-square estimators in the time-series model. In the time-series case the term to be minimised in maximum-likelihood estimation is

$$KU = \frac{\boldsymbol{u}'\mathbf{V}^{-1}\boldsymbol{u}}{(1-\rho^2)^{1/n}} . \tag{30}$$

As n tends to infinity it is clearly seen that

$$\lim_{n\to\infty}(1-\rho^2)^{1/n} = 1 , \tag{31}$$

so that the maximum-likelihood estimator is asymptotically equivalent to minimising $\boldsymbol{u}'\mathbf{V}^{-1}\boldsymbol{u}$. In any finite sample the minimum chi-square and maximum-likelihood estimates will differ numerically because of $(1-\rho^2)^{1/n}$; however, their asymptotic properties are the same.

It is possible that in the spatial maximum-likelihood model K is such a weak penalty function that omission of it makes very little difference to the estimates, so that the simpler nonlinear least-squares method might be used in place of the maximum-likelihood method. This is not the case, as may be shown by the following illustration.

The eigenvalues of **M** were extracted for spatial autoregression on the lattice of 49 states of the USA (48 continental states plus the District of

Columbia). **M** was scaled to have row sums equal to unity so that each state value was a linear function of the average contiguous value. This gives real eigenvalues of modulus $\leqslant 1 \cdot 0$, with ρ in the range $-1 \cdot 0 < \rho < +1 \cdot 0$ permitted. K and $|\mathbf{P}|$ were then evaluated for $\rho = -1 \cdot 0$ to $+1 \cdot 0$ in steps of $0 \cdot 01$, to trace out the form and magnitude of the penalty function. These are charted in figure 1. The form of K, growing greater than $1 \cdot 0$ as ρ moves to either $+1 \cdot 0$ or $-1 \cdot 0$ penalises least-squares solutions that give large absolute $\hat{\rho}$ values. The very high penalty values as ρ approaches $\pm 1 \cdot 0$ (there is a singularity at $+1 \cdot 0$) restrict the maximum-likelihood estimates to within the feasible range; there is no such constraint in the least-squares estimator, with K set to unity, and values of ρ greater than

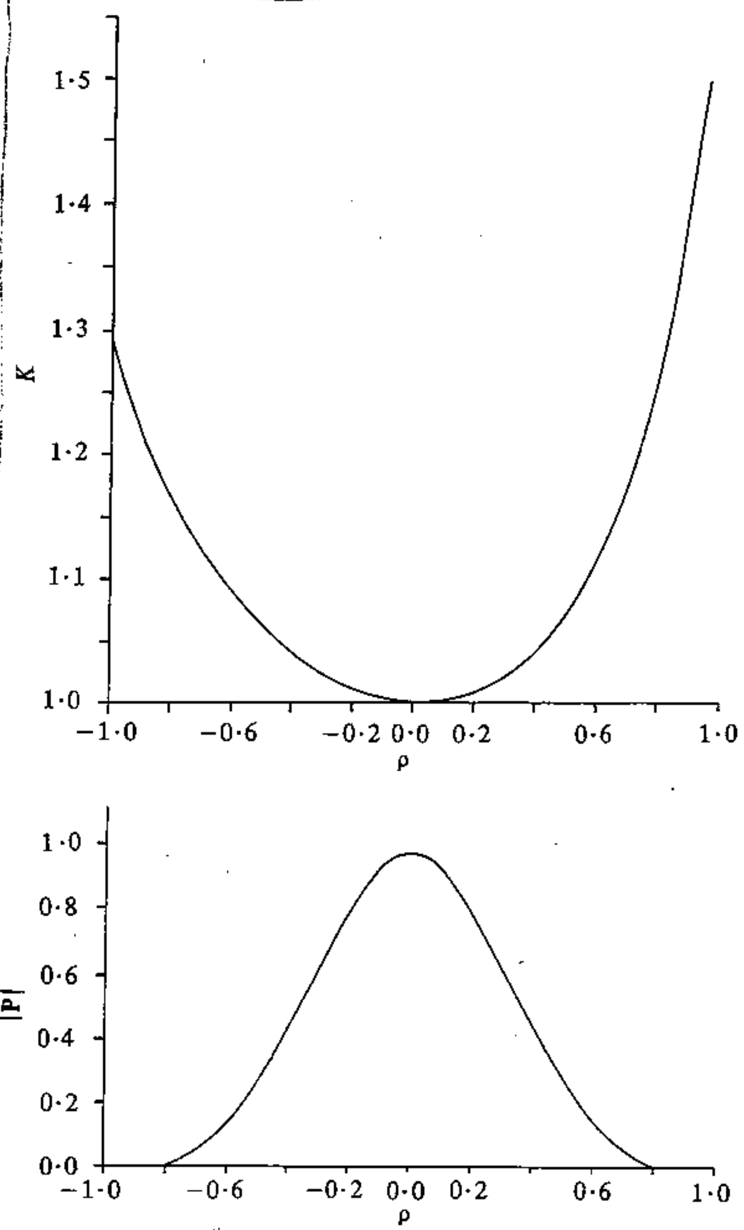

Figure 1. The Jacobian penalty term K and the Jacobian $|\mathbf{P}|$ for various ρ values in the lattice for the United States.

$1 \cdot 0$ are frequently generated as least-squares estimates. This structure of KU suggests that in general the least-squares estimator will be biased towards overestimating the absolute value of ρ.

The spatial maximum-likelihood model is in many ways similar to the simultaneous equation model of econometrics. As in the simultaneous-equation model, least-squares procedures are invalid because of the presence of a nonvanishing Jacobian element (though in the simultaneous-systems case there are adaptations of least-squares procedures, such as two- and three-stage least-squares, that take account of the Jacobian). The Jacobian in the spatial-series model also plays a role directly analogous to the Jacobian term containing the matrix of endogenous coefficients in simultaneous-equation systems. Thus in Haavelmo's (1947) consumption model for example, the Jacobian term $|1-\alpha|^n$, where α is the endogenous coefficient, may be shown to be a penalty function, that compensates for the tendency of the minimum-sum-of-squares estimator to overestimate the absolute value of α (Zellner, 1971), as noted above for ρ in the spatial model.

7 Empirical application

In this section the maximum-likelihood model is illustrated by application to a regression equation in a paper by Hanna (1966). Hanna's paper is an econometric study of the effects of regional differences in sales taxes and transport charges across the United States on the pattern of car ownership or 'automobile consumption'. The models were set up in the form of the effects that taxes, etc. on new cars would have on the value of second-hand cars, and hence on the relative frequency of older cars from state to state.

The particular equation investigated was

$$y = \alpha + \beta x + u \,, \tag{32}$$

where y is the vector of 49 observations on the average value of 1955–1959 model cars in January 1960 for each state, and x is the vector of observations on new-car price differentials attributable to transport charges and sales taxes. Both variables are expressed in dollars. β is clearly expected to be >0 but probably <1. The least-squares estimates are given in table 1. These double-precision estimates differ slightly, but

Table 1. Ordinary least-squares estimates.

	α	β	
standard errors	27·20	0·1947	$\hat{\sigma}_u^2 = 3181 \cdot 97$
t values	52·76	3·5245	$r^2 = 0 \cdot 250$

$$y = 1435 \cdot 97 + 0 \cdot 6864x + u$$

not significantly, from those given by Hanna (1966). The slope coefficient, $\hat{\beta}$, is 0·686, positive and <1, but significantly different from zero.

The least-squares regression residuals, $\hat{\boldsymbol{u}}$, were then tested for spatial autocorrelation between adjacent states by the statistic

$$s = \frac{\hat{\boldsymbol{u}}'\mathbf{W}\hat{\boldsymbol{u}}}{\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}} , \tag{33}$$

where $\mathbf{W}$ is the $n \times n$ matrix of contiguities, where entry $w_{ij} = 1$ if regions i and j are contiguous, and zero otherwise. Since $\hat{\boldsymbol{u}} = \mathbf{G}'\boldsymbol{u}$ where $\mathbf{G}$ is the idempotent matrix $\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ (Johnston, 1972), the statistic s is distributed as

$$s = \frac{\boldsymbol{u}'\mathbf{G}'\mathbf{W}\mathbf{G}\boldsymbol{u}}{\boldsymbol{u}'\mathbf{G}\boldsymbol{u}} . \tag{34}$$

Under independence of the errors $\boldsymbol{u}$ this may be shown (Hepple, 1974a) to be asymptotically normally distributed with mean value

$$\mathrm{E}(s) = \frac{\operatorname{tr}\mathbf{G}'\mathbf{W}}{n-k} , \tag{35}$$

and variance

$$\operatorname{var}(s) = 2\left[\frac{(n-k)\operatorname{tr}(\mathbf{G}'\mathbf{W})^2-(\operatorname{tr}\mathbf{G}'\mathbf{W})^2}{(n-k)^2(n-k+2)}\right] . \tag{36}$$

For Hanna's model, $s = 2{\cdot}763$. Under independence the expected value is $-0{\cdot}129$, with variance 0·148, thus giving a standard normal deviate of 7·506, which is highly significantly different from zero. This inference was also confirmed by the application of exact, small-sample tests. Since the tests indicate highly significant spatial dependence in the errors, the regression model with spatially autoregressive errors was estimated:

$$\begin{aligned} y &= \alpha+\beta x+\boldsymbol{u} , \\ \boldsymbol{u} &= \rho\mathbf{M}\boldsymbol{u}+e , \end{aligned} \tag{37}$$

with $e \sim N(0, \sigma^2)$. The eigenvalues of the 49×49 nonsymmetric matrix $\mathbf{M}$ were first computed, with $\mathbf{M}$ scaled so that row sums were unity, as in section 6, and the coefficients of equation (37) were then estimated by Powell's conjugate gradient algorithm. The algorithm converged to a maximum after eight iterations and 129 function evaluations. The estimates are given in table 2, together with asymptotic standard errors and significance tests. Incorporation of the dramatically large $\hat{\rho}$ value of 0·816 has significantly reduced the slope-coefficient estimate. It is now close to, and not significantly different from, zero. It is possible, though unlikely, that the high autoregressive coefficient is the result of a local rather than a global maximum in the likelihood function. To check this separate maximisations were started from initial ρ values of 0·0, 0·4, and 0·8, and each converged to the same estimates as in table 2, so that the estimates represent a global maximum.

In order to assess whether inclusion of the Jacobian term in the maximum-likelihood estimation makes a real difference in an empirical situation, the model was also estimated leaving out this term, that is by minimum chi-square, by minimising $\boldsymbol{u}'\mathbf{V}^{-1}\boldsymbol{u}$ and assuming the Jacobian term to be one. Simultaneous estimation of α, β, and ρ by nonlinear least squares, again using Powell's algorithm, produced the following minimum chi-square estimates:

$$y = 1579{\cdot}11 + 0{\cdot}0534x + u,$$

$$u = 0{\cdot}9785\mathbf{M}u + e\,, \qquad \hat{\sigma}_e^2 = 966{\cdot}97\,,$$

with convergence after 62 function evaluations. The estimated $\hat{\rho}$ value is considerably larger than in table 2, 0·979, thus indicating that inclusion of the Jacobian penalty term works in the way suggested in section 6. The $\hat{\beta}$ estimate is still lower than in the maximum-likelihood estimation, at 0·053. Spatial versions of the Durbin procedure and the Cochrane–Orcutt method, both asymptotically equivalent to minimising $\boldsymbol{u}'\mathbf{V}^{-1}\boldsymbol{u}$, produced very similar results in this application.

The maximum-likelihood results show the importance of incorporating spatial dependence into the estimation of econometric relationships. In this particular case, incorporation of spatial dependence results in rejection of the inference that new-car price differentials attributable to sales taxes and transport charges significantly affect the value of used cars. Comparison of the maximum-likelihood and minimum chi-square estimates also indicates that the simpler but inconsistent minimum chi-square procedures lead to substantially different numerical results to those by consistent maximum-likelihood estimation. Inclusion of the Jacobian term is therefore important.

Table 2. Maximum-likelihood estimates.

	Asymptotic standard error	Asymptotic t value
$\hat{\alpha}$	31·791	48·073
$\hat{\beta}$	0·121	0·820
$\hat{\rho}$	0·074	11·035

$$y = 1528{\cdot}17 + 0{\cdot}0992x + u\,,$$

$$u = 0{\cdot}8164\mathbf{M}u + e\,, \qquad \hat{\sigma}_e^2 = 1044{\cdot}74\,.$$

8 Concluding remarks

The maximum-likelihood spatial estimator in this paper suffers from the general drawback shared by the maximum-likelihood and minimum chi-square time-series estimators: the desirable statistical properties are all asymptotic. A further, highly theoretical qualification is discussed in the appendix. In small, finite samples maximum-likelihood estimates need not be unbiased and normally distributed. Hildreth has shown, for the time-series model, that it is likely that the distribution of $\hat{\beta}$ in autoregressive models rapidly approaches normality as the sample size increases, but that larger samples will be needed for normality of $\hat{\rho}$ and $\hat{\sigma}^2$ (Hildreth, 1969). Several Monte Carlo simulation studies of small-sample properties have been made for the time-series estimators (for example Griliches and Rao, 1969; Hendry and Trevedi, 1970). Some very preliminary small-sample Monte Carlo studies of the maximum-likelihood spatial model (Hepple, 1974a) suggest that there is some downward bias of $\hat{\rho}$, but that on both average bias and mean-square-error criteria, the maximum-likelihood estimates are superior to direct minimisation of the sum of squares, i.e. when the Jacobian term is ignored.

It may prove possible to construct more accurate approximations to the sampling distributions than by assuming normality. In the time-series case Hildreth has suggested the possibility of a beta-distribution for $\hat{\rho}$. An alternative approach is to stop focussing purely on the maximum of the likelihood, and study the whole likelihood function, or recast the estimation in Bayesian terms (Zellner, 1971). Suitable diffuse prior distributions still allow the likelihood to dominate the posterior distribution, and the method allows exact finite-sample inference, though not of course in the same terms as the classical inference of maximum likelihood. A Bayesian formulation and application of the present model will be presented in another paper.

The maximum-likelihood estimator does however extend readily to simultaneous-equation systems with spatially autocorrelated errors (Hepple, 1974a), whereas it is difficult, on purely computational grounds, to envisage a full Bayesian analysis of such models or of large multiple regressions, though marginal distributions may be obtained.

The maximum-likelihood model that has been outlined in this paper is only one of a family of possible models that build spatial dependence into econometric estimation. It is important to also explore moving average, Box–Jenkins, and other specifications of the error structure. Alternatively, the spatial autocorrelation in the residuals may be the result of misspecification of the actual regression structure, this requiring the introduction of spatially lagged X variables or lagged dependent variables, rather than directly modelling dependence in the errors. However the maximum-likelihood model presented here shows the main features and problems that arise in econometric estimation with spatial data, together with the empirical importance of incorporating spatial dependence into the estimation.

Acknowledgement. The author is indebted to Professor Peter Whittle for his helpful comments on this model.

References

Box, G. E. P., Jenkins, G. M., 1970, *Time Series Analysis, Forecasting and Control* (Holden-Day, San Francisco).

Cliff, A. D., Ord, J. K., 1973, *Spatial Autocorrelation* (Pion, London).

Cochrane, D., Orcutt, G. H., 1949, "Application of least-squares regressions to relationships containing autocorrelated error-terms", *Journal of the American Statistical Association*, **44**, 32-61.

Cooper, J. P., 1972, "Asymptotic covariance matrix of procedures for linear regression in the presence of first-order serially correlated disturbances", *Econometrica*, **40**, 305-310.

Cramer, H., 1945, *Mathematical Methods of Statistics* (Princeton University Press, Princeton).

Dhrymes, P. J., 1971, *Distributed Lags: Problems of Estimation and Formulation* (Oliver and Boyd, Edinburgh).

Durbin, J., 1960, "Estimation of parameters in time-series regression models", *Journal of the Royal Statistical Society*, 22B, 139-153.

Fletcher, R., 1970, *Optimization* (Academic Press, London).

Godfrey, L., 1971, "The Phillips curve: incomes policy and trade union effects", in *The Current Inflation*, Ed. H. G. Johnson and A. R. Nobay (Macmillan, London), pages 99-124.

Goldfeld, S. M., Quandt, R. E., 1972, *Nonlinear Methods in Econometrics* (North-Holland, Amsterdam).

Griliches, Z., Rao, P., 1969, "Small-sample properties of several two-stage regression methods in the context of autocorrelated errors", *Journal of the American Statistical Association*, **64**, 253-272.

Haavelmo, T., 1947, "Methods of measuring the marginal propensity to consume", *Journal of the American Statistical Association*, 42, 105-122.

Hanna, F. A., 1966, "Effects of regional differences in taxes and transportation charges on automobile consumption", in *Papers on Regional Statistical Studies*, Ed. S. Ostry and T. K. Rhymes (Toronto University Press, Toronto), pp.199-223.

Hendry, D. F., 1971, "Maximum likelihood estimation of systems of simultaneous regression equations with errors generated by a vector autoregressive process", *International Economic Review*, **12**, 257-272.

Hendry, D. F., Trevedi, P. K., 1970, "Maximum likelihood estimation of difference equations with moving average errors: a simulation study", Discussion Paper No. 7, London School of Economics and Political Science, Department of Economics.

Hepple, L. W., 1974a, *Econometric Estimation and Model-Building with Spatial Series*, Ph. D. Thesis, University of Cambridge.

Hepple, L. W., 1974b, "The impact of stochastic process theory upon spatial analysis in human geography", in *Progress in Geography 6*, Eds C. Board, R. J. Chorley, P. Haggett and D. R. Stoddart (Arnold, London), pp.89-142.

Hepple, L. W., in preparation, *Econometric Estimation with Spatial Series* (Pion, London).

Hildreth, C., 1969, "Asymptotic distribution of maximum likelihood estimators in a linear model with autoregressive disturbances", *Annals of Mathematical Statistics*, **40**, 583-594.

Hildreth, C., Lu, J. Y., 1960, "Demand relations with autocorrelated disturbances", Research Bulletin 276, Agricultural Experiment Station, Michigan State University, East Lansing, Michigan, USA.

Himmelblau, D. M., 1972, *Applied Nonlinear Programming* (McGraw-Hill, New York).
Hoeffding, W., Robbins, H., 1948, "The central limit theorem for dependent variables", *Duke Mathematical Journal*, 15, 773-780.
Johnston, J., 1972, *Econometric Methods*, 2nd edition (McGraw-Hill, New York).
Lee, M. L., 1970, "A model of the distribution of federal expenditures among States", *Journal of the American Statistical Association*, 65, 136-149.
Mead, W. R., 1967, "A mathematical model for the estimation of interplant competition", *Biometrics*, 23, 189-205.
Powell, M. J. D., 1965, "An efficient method for finding the minimum of a function of several variables without calculating derivatives", *Computer Journal*, 7, 155-162.
Steinnes, D. N., Fisher, W. D., 1974, "An econometric model of intraurban location", *Journal of Regional Science*, 14, 65-80.
Theil, H., 1971, *Principles of Econometrics* (Wiley, New York).
Wallis, K. F., 1971, "Wages, prices and incomes policies: some comments", *Economica*, 38, 304-310.
Whittle, P., 1954, "On stationary processes in the plane", *Biometrika*, 41, 434-449.
Wilkinson, J. H., 1965, *The Algebraic Eigenvalue Problem* (Clarendon Press, Oxford).
Zellner, A., 1971, *An Introduction to Bayesian Inference in Econometrics* (Wiley, New York).

Appendix

Properties of maximum-likelihood estimators

The classical proofs of the properties of maximum-likelihood estimators, the properties of consistency and asymptotic multivariate normality, were derived on the basis of independently distributed observations (Cramer, 1945, pages 500-504). Intuitively, these properties will also hold for temporal and spatial models where the observations are no longer independent, provided that the structure of the generating process is not peculiar (for example if matrix **M** were broken into disjoint 'islands'). To be completely rigorous, however, the properties of maximum-likelihood estimators need to be proved again for the temporal and spatial models. This theoretical problem has been largely ignored by econometricians and others studying time-series regression models (for example Cochrane and Orcutt, 1949; Cooper, 1972; Hendry and Trevedi, 1970), but in recent years a few papers have proved the maximum-likelihood properties for the regression model with temporally autoregressive errors (Dhrymes, 1971; Hildreth, 1969). Their proofs, based on the theory of m-dependent variables (Hoeffding and Robbins, 1948), are not easily extended to the spatial maximum-likelihood models discussed here. Nor have proofs yet been constructed for simultaneous equation models with temporally autocorrelated errors (Hendry, 1971). As in the time-series literature, it would be pedantic to delay exploration and application of the spatial maximum-likelihood models for want of this detailed proof. Equally, it would be valuable to have such a proof.