Available at
www.ElsevierMathematics.com
POWERED BY SCIENCE DIRECT®
ELSEVIER
Computational Statistics & Data Analysis 45 (2004) 179–196
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Chebyshev approximation of log-determinants of spatial weight matrices

R. Kelley Pace[a,*], James P. LeSage[b]

[a] *Department of Finance, E.J. Ourso College of Business Administration, Louisiana State University, Baton Rouge, LA 70803-6308, USA*
[b] *Department of Economics, University of Toledo, Toledo, OH 43606, USA*

Received 1 June 2001; received in revised form 1 September 2002

Abstract

To cope with the increased sample sizes stemming from geocoding and other technological innovations, this paper introduces an $O(n)$ approximation to the log-determinant term required for likelihood-based estimation of spatial autoregressive models. It takes as a point of departure Martin's (1993) Taylor series approximation based on traces of powers of the spatial weight matrix. Using a Chebyshev approximation along with techniques to efficiently compute the initial matrix power traces results in an extremely fast approximation along with bounds on the true value of the log-determinant. Using this approach, it takes less than a second to compute the approximate log-determinant of an $890,091 \times 890,091$ matrix. This represents a tremendous increase in speed relative to exact computation that should allow researchers to explore much larger problems and facilitate spatial specification searches.

Keywords: Spatial statistics; Spatial autoregression; Maximum likelihood; Sparse matrices; Log-determinants; Chebyshev matrix determinant approximations

1. Introduction

While technology increases the ability to analyze data, it also increases the ability to collect and disseminate data. Arguably, the collection and dissemination of spatial data have grown faster than the ability to analyze it. For example, the publicly available Home Mortgage Disclosure Act (HMDA) data now has over 100 million mortgage

* Corresponding author. Tel.: +225-578-6256; fax: +225-578-9066.
E-mail addresses: kelley@pace.am (R.K. Pace), jlesage@spatial-econometrics.com (J.P. LeSage).
URL: http://www.spatial-statistics.com

0167-9473/$ - see front matter
PII: S0167-9473(02)00321-3

loan application records. The year 2000 Census provides data on over 200 thousand block groups and millions of blocks. In the context of spatial data mining, Francica (2000, p. 21) reports firms using spatial databases with over a terabyte of data. All of these data have an important spatial component. For example, housing prices act as a sensitive barometer for many social phenomenon such as crime, congestion, job opportunities, and demographics.

Many of these variables display substantial amounts of spatial dependence making aspatial regression methods inappropriate. Ideally, researchers could employ likelihood-based estimation methods (maximum likelihood or Bayesian) to analyze these large spatial data samples using all relevant observations and variables as opposed to deriving some ad hoc procedure for computational feasibility. Likelihood-based methods draw upon a rich corpus of statistical knowledge concerning inference and can coherently handle both linear and non-linear model specifications. Note, computational advances of the type described here benefit both maximum likelihood and Bayesian estimation methods because of their common reliance on the likelihood function.

For the simultaneous spatial autoregressions commonly employed in the economics and geography literature, the normal density profile or concentrated log-likelihood function equals $C + \ln|I - \alpha D| - (n/2)\ln(SSE(\alpha))$, where C represents an additive constant, n represents the number of observations, α represents a scalar parameter measuring the strength of spatial dependence, SSE represents the sum-of-squared-errors, D represents an $n \times n$ spatial weight matrix, and $\ln|I - \alpha D|$ represents a log-determinant term (Pace and Barry, 1997, p. 235). Evaluating the log-determinant term of the $n \times n$ matrix D constitutes the binding computational constraint on spatial estimation. Computing the log-determinant requires $O(n^3)$ operations when done in the most straightforward manner. The log-determinant term ensures that the density and hence the likelihood remains proper (integrates to one). It arises not only for the Gaussian density but for other continuous proper densities such as the Student t, Wishart, and Laplace. Bayesian estimation methods, with their focus on the underlying likelihood function, also require computation of the log-determinant. For example, Markov Chain Monte Carlo estimation of Bayesian spatial autoregressive models proposed by LeSage (1997, 2000) uses the log-determinant in Metropolis sampling to evaluate the conditional distribution of α.

In response to the computational challenge posed by the log-determinant, at least four strategies exist. First, one could randomly sample the data, but this may destroy spatial relationships that are the focus of interest. Since sample information on spatial dependence often improves prediction and the precision of inference, the introduction of a higher average distance between observations through sampling may reduce the value of the regression exercise. A second approach involves tessellation of the data into smaller areas that can be fit using separate models for each area. If the analysis seeks to provide an overall relation or summary of the relationship among variables in the model, tessellation may defeat the purpose here as well. In addition, the logistics of managing many separate models may prove daunting. A third approach involves the use of alternative estimators, but these often encounter difficulties in drawing inferences. The instrumental variable approach of Anselin (1988, pp. 81–90) and the two-stage least-squares approach of Kelejian and Prucha (1998) exemplify this strategy. The fourth strategy is to directly attack the computational difficulties confronting

likelihood estimation. The spectral method of Whittle (1954), the Taylor series approach of Martin (1993), the eigenvalue-based approach of Griffith and Sone (1995), the direct sparse matrix approach of Pace and Barry (1997), the Monte Carlo approach of Barry and Pace (1999), the graph theory approach of Pace and Zou (2000), and the characteristic polynomial approach of Smirnov and Anselin (2001) represent examples of this strategy.

Besides generality, directly addressing the computational problems posed by the log-determinant allows partitioning the estimation problem into two parts. The first part involves approximation (or exact computation) of the log-determinant and the second part involves statistical estimation based on the log-likelihood. By construction, errors from the numerical approximation should remain independent of the random variables in the statistical model. This facilitates the development and testing of approaches for calculating the log-determinant.

The separate computation of an exact or approximate log-determinant function assists in likelihood ratio tests, a powerful method for conducting inference. This approach to testing involves solving multiple maximum problems involving the likelihood for different explanatory variable configurations of the model. As the submodels use the same log-determinant function, the computational advantages of computing this once and reusing it become clear.

This paper reexamines and builds upon the approximation proposed by Martin (1993) who used a Taylor series to approximate the trace of the matrix logarithm (which equals the log-determinant). In Martin's approach, the Taylor series is used to approximate the function $\sum_{i=1}^{n} \ln(1-\alpha\lambda_i)$, where λ_i represents the ith eigenvalue that lies in the interval $[-1,1]$ and α represents a scalar parameter from the interval $(-1,1)$.[1] Unfortunately, Taylor series do not usually perform well over intervals, a point summarized by Muller (1997, p. 29), "...Taylor expansions only give local (i.e., around one value) approximations, and should not be used for global (i.e., over an interval) approximations." Muller proceeds to give an example where the Taylor series performs more than one million times worse than the best polynomial approximation.

Chebyshev orthogonal polynomials serve as the most common technique for functional approximation over a closed interval. For this problem, switching to a Chebyshev approximation provides a large performance improvement over Martin's Taylor series approximation. Amazingly, the estimated autoregressive parameters for a 57,647 observation Census tract example used here were within 0.02 of the exact values, but these estimates were obtained over 600 times faster than using an exact method. This ratio of improved computational speed increases with n. Computing the Chebyshev quadratic approximation took under a second for a data sample based on 890,091 US

[1] A priori the eigenvalues of D fall into the interval $[-1,1]$. If one knew the minimum eigenvalue $\lambda_{\min}$ ($\lambda_{\min} > -1$), this would aid the approximation by reducing the domain of the function. In this case, α would fall in the interval $(\lambda_{\min}^{-1}, 1)$. However, since finding the minimum eigenvalue requires more computation than the entire log-determinant approximation and since autoregressive parameters rarely take on values less than -1 in applied practice, we require the approximation to work for eigenvalues in $[-1,1]$ rather than in the shorter eigenvalue interval $[\lambda_{\min}, 1]$. See Smirnov and Anselin (2001) for a discussion of the unattractiveness of $\alpha < -1$ in empirical problems.

Census block observations. Computing the exact log-determinant for a dataset of this size would have taken far more time and memory.

Often approximations work, but provide little information concerning their accuracy. To address this, Martin's Taylor series approximation, which has known biases, provides useful upper and lower bounds on the accuracy of the Chebyshev log-determinant approximation.

Section 2 discusses the mathematics and techniques behind the approximation and the bounds. Section 3 provides applied illustrations of the techniques from Section 2, while Section 4 concludes the paper.

2. Approximation of the log-determinant

This section presents the theoretical background behind the improved log-determinant approximations and bounds. Section 2.1 defines various spatial weight matrices and discusses their properties. Section 2.2 introduces Taylor series approximations of the log-determinant to provide a setting for the Chebyshev approximation of the log-determinant contained in Section 2.3. Section 2.4 proposes use of Taylor series to determine bounds applicable to the log-determinant approximation, and Section 2.5 examines efficient approaches for computing the traces required by the approximation. Finally, Section 2.6 gives the spatial regressive autoregressive model employed in the applied illustrations of Section 3.

2.1. Spatial weight matrix properties

The properties of the spatial weight matrix used affect the approximation of the log-determinant. The techniques proposed herein assume the spatial weight matrix is a real, non-negative, $n \times n$, sparse matrix with the maximum eigenvalue equal to 1, minimum eigenvalue greater than or equal to -1, and a trace equal to 0. These are not restrictive assumptions as they apply to a great many commonly used spatial weight matrices. For convenience, let D denote any spatial weight matrix with these properties. For the techniques discussed in this paper we assume that the spatial weight matrix is symmetric or similar to a symmetric matrix. Fortunately, two common methods for constructing spatial weight matrices, nearest neighbors and Delaunay triangles, can meet this requirement. Note, one can partition the spatial estimation problem into a part involving the computation of the log-determinant term and a part involving the statistical computation of the parameter estimates. This partitioning permits use of a constructed symmetric version of the weight matrix when computing the log-determinant and the use of a similar row-stochastic version of the weight matrix to carry out the statistical computations.

The use of the m nearest neighbors provides one way of constructing a spatial weight matrix, where $N_1, N_2, \ldots, N_m$ represent a sequence of m individual nearest neighbor weight matrices. Each of these individual neighbor matrices contains a single 1 in each row with all other entries equal to zero. The first individual neighbor matrix contains the very nearest neighbor to each observation, the second individual neighbor

matrix contains the second nearest neighbor to each observation, and so forth. Let D_N represent a weighted average of the individual neighbor matrices. Specifically, let $D_N = w_1N_1 + w_2N_2 + \cdots + w_mN_m$ where $w_1 + w_2 + \cdots + w_m = 1$ and $w_1, w_2, \ldots, w_m \geqslant 0$. By construction, D_N is row-stochastic (i.e., $D_N\imath = \imath$, where $\imath$ is a column vector of ones) and this implies it has a maximum eigenvalue equal to 1 and a minimum eigenvalue greater than or equal to -1. Naturally, $N_1, N_2, \ldots, N_m$ are asymmetric and hence D_N is usually asymmetric.

We can however transform D_N to produce a related symmetric matrix D_S. Define r_i as the row-sum of the ith row of $D_N + D_N'$. Since N has m positive entries in each row, $r_i > 0$. Let R be an $n \times n$ diagonal matrix with $R_{ii} = r_i$ and let $D_R = R^{-1}(D_N + D_N')$. Hence, D_R is also row-stochastic (i.e., $D_R\imath = \imath$). Let $D_S = R^{-1/2}(D_N + D_N')R^{-1/2}$. While D_R is usually asymmetric, D_S is symmetric. As Ord (1975, p. 125) pointed out, D_R and D_S have the same eigenvalues (D_R, D_S are similar). Moreover, the eigenvalues for both matrices are real, since symmetric real matrices like D_S have real eigenvalues (Marcus and Minc, 1992, p. 64). Note, the symmetric transformation means D_S, D_R now may have a variable number of neighbors to each observation.

The use of Delaunay triangles provides another way of constructing a spatial weight matrix. If a leg of a Delaunay triangle connects observation i and j, set the elements in the unweighted spatial weight matrix $D_{ij}^u = 1$. This will naturally result in a symmetric matrix. Upon setting all elements associated with legs of the triangles to 1, one can reweight the matrix as with nearest neighbors to create a row-stochastic matrix and a symmetric matrix similar to the row-stochastic matrix. As noted earlier, one can use the symmetric version of the weight matrix for log-determinant calculations while relying on the row-stochastic version for the statistical aspects of estimation.

2.2. Taylor series approximation

The Taylor series approximation of the matrix logarithm of $(I - \alpha D)$ in (1) has a simple form (Golub and Van Loan, 1996, p. 566).

$$\ln(I - \alpha D) = -\sum_{j=1}^{\infty} \frac{\alpha^j D^j}{j}. \tag{1}$$

As Martin (1993) noted, the log of the determinant of $(I - \alpha D)$ equals the trace of the matrix logarithm, which equals the sum of $\ln(1 - \alpha\lambda_i)$ for $i = 1, \ldots, n$ where λ_i represents the ith eigenvalue of D and due to the linearity of the trace function the evaluation of the log-determinant across all α does not require recomputing the moments but only the reweighting of some scalars (the traces of the moments). To make this implementable, the series in (2) terminates after q terms

$$\ln|I - \alpha D| \simeq -\sum_{j=1}^{q} \frac{\alpha^j \, tr(D^j)}{j}. \tag{2}$$

Taylor series approximations typically do well around the point of expansion x (in this case $x = 0$), but their performance degrades for arguments away from the expansion point.

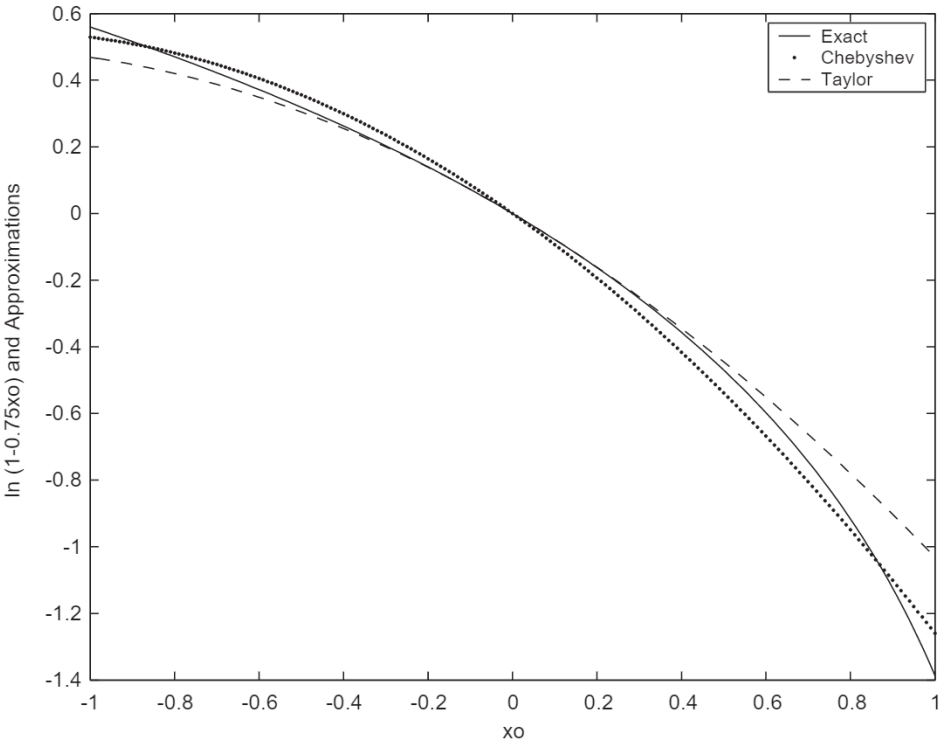

Fig. 1. Quadratic ln approximation.

2.3. Chebyshev approximation

In contrast to the Taylor series local approximation, Chebyshev approximations strive to provide uniform performance over some interval. Since the weight matrix eigenvalues fall in $[-1, 1]$, an interval approximation seems appropriate. Fig. 1 shows the quadratic Chebyshev and Taylor series approximations for the function where x varies from -1 to 1 and $\alpha = 0.75$. This figure reveals two features of both approximation techniques. First, the Taylor series approximation does better than the Chebyshev approximation for values of x nearer 0, but far worse for extreme x. Second, the Chebyshev approximation crosses the 0 error line 3 times (the degree of the polynomial used plus 1) and its error alternates in sign. In contrast, the Taylor series approximation crosses the 0 error line once at $x = 0$.

A log-determinant approximation sums the approximated values of $\ln(1 - \alpha\lambda_i)$ over all i.[2] Sometimes the approximation to $\ln(1 - \alpha\lambda_i)$ overpredicts its true value

[2] Since we employ symmetric and real weight matrices, these are real normal matrices (Marcus and Minc, 1992, p. 62). Consequently, all the eigenvalues are real and the matrix problem reduces to the scalar problem (Toh and Trefethen, 1998, p. 401). For non-symmetric matrices the Chebyshev approximation becomes more difficult.

(negative errors) and sometimes it underpredicts its true value (positive errors). The exact number of positive and negative errors depends upon the distribution of λ. Summing over approximation errors of different signs can materially decrease the overall error. In conjunction with these differences in sign changes, note that the Taylor series produces different error magnitudes for positive versus negative λ_i, whereas the Chebyshev tends to produce errors of more equal maximum magnitude. This potential error canceling feature of the log-determinant favors the Chebyshev over the Taylor series approximation.

Let q represent the highest power of the approximating polynomial which thus has $q+1$ coefficients. The Chebyshev approximation in (3) uses a linear combination (scalars $c_1(\alpha),\ldots,c_{q+1}(\alpha)$) of matrix polynomials

$$\ln|I-\alpha D| \simeq \sum_{j=1}^{q+1} c_j(\alpha)\, tr(T_{j-1}(D)) - \frac{1}{2}\, c_1(\alpha), \tag{3}$$

where $T_0(D)=I$, $T_1(D)=D$, $T_2(D)=2D^2-I$, $T_3(D)=4D^3-3D$, and $T_4(D)=8D^4-8D^2+I$ (Press et al., 1996, p. 184). This paper only employs $T_0(D)\ldots T_4(D)$, but additional terms in the sequence use the relation $T_{k+1}(D)=2DT_k(D)-T_{k-1}(D)$.

As illustrated by Fig. 1, the Chebyshev approximation error equals 0 at $q+1$ known locations over the interval $[-1,1]$. Press et al. (1996, p. 185) use the $q+1$ points of zero error to derive the expression in (4) for the coefficients $c_1(\alpha),\ldots,c_{q+1}(\alpha)$. Specifically, (4) is their equation 5.86 with substitution of the specific function $\ln(1-\alpha x)$ for their more general $f(x)$.

$$c_j(\alpha)=\left(\frac{2}{q+1}\right)\sum_{k=1}^{q+1}\ln\left[1-\alpha\cos\left(\frac{\pi(k-1/2)}{q+1}\right)\right]\cos\left(\frac{\pi(j-1)(k-1/2)}{q+1}\right). \tag{4}$$

The coefficients in (4) depend upon α, so as α becomes smaller the function becomes easier to approximate, and the coefficients adapt to the new function. After computing the moments, the approximation only involves operations on scalars which do not require much computation time or storage.

Fig. 2 shows the quadratic Chebyshev approximation, the exact log-determinant, and Taylor series bounds for a data set with 57,647 observations using six nearest neighbors that are weighted equally. For small and moderate values of α the curves run together, but they begin to separate at higher levels of α. Even for the larger values of α, the Chebyshev and exact log-determinant curves exhibit a very similar shape while the Taylor series curves diverge. Fig. 3 shows the a differenced version of the same functions.

Any low-order approximation can encounter difficulty. In a univariate problem, one might evaluate the approximation at equally spaced points, and the average of the errors will be low by construction. However, evaluation of a matrix does not necessarily resemble the univariate problem since the eigenvalues may not lie uniformly in the interval. For the Chebyshev, if the eigenvalues of the weight matrix concentrate in the region of the approximation with the least accuracy, naturally the approximation may not perform well. As the order of approximation changes, these points will also change. Thus, a second-order approximation could perform better than a fourth-order

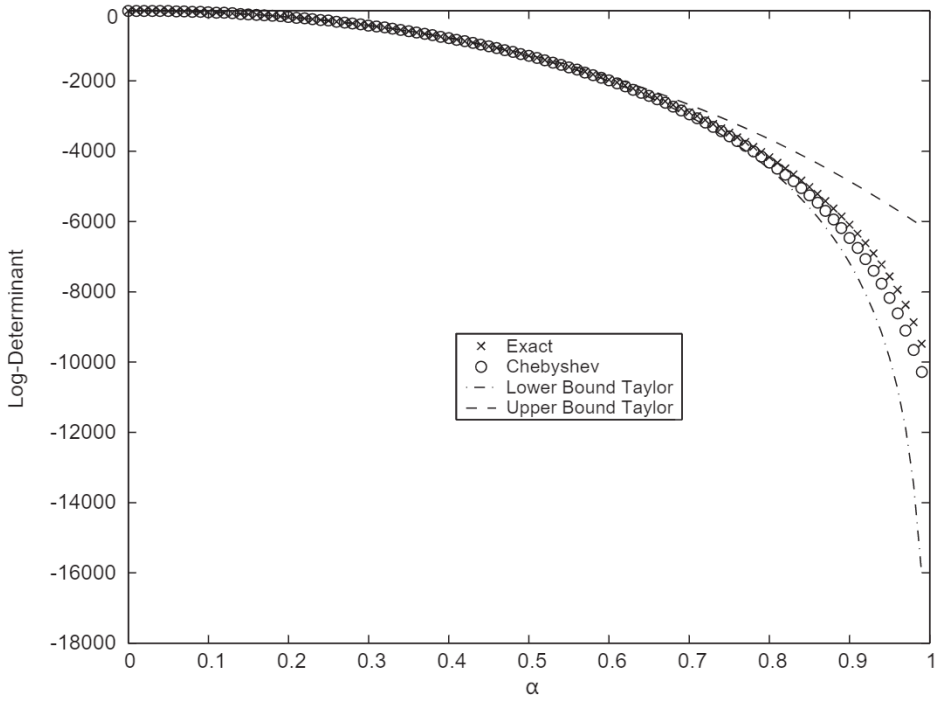

Fig. 2. Exact and approximate log-determinants for US Census tract locations.

approximation for a particular matrix, depending upon the location of the eigenvalues. Unfortunately, calculating the eigenvalues of the matrix becomes prohibitively expensive for large matrices (this would also obviate any need to approximate), and so one should select an approximation which performs well ex ante. The Chebyshev provides one of the better means of minimizing such ex-ante error.

2.4. Log-determinant bounds using Taylor series

While Taylor series may not prove ideal for approximation of a function over an interval, the series for $\log(1-\alpha x)$ has predictable biases. Assuming that we have computed q moments, Eq. (5) defines the upper and the lower log-determinant bounds.

$$\ln|I-\alpha D|_U = -\sum_{j=1}^{q}\frac{\alpha^j\, tr(D^j)}{j} + \left(-\sum_{j=q+1}^{\infty}\frac{\alpha^j B(j)_L}{j}\right),$$

$$\ln|I-\alpha D|_L = -\sum_{j=1}^{q}\frac{\alpha^j\, tr(D^j)}{j} + \left(-\sum_{j=q+1}^{\infty}\frac{\alpha^j B(j)_U}{j}\right). \tag{5}$$

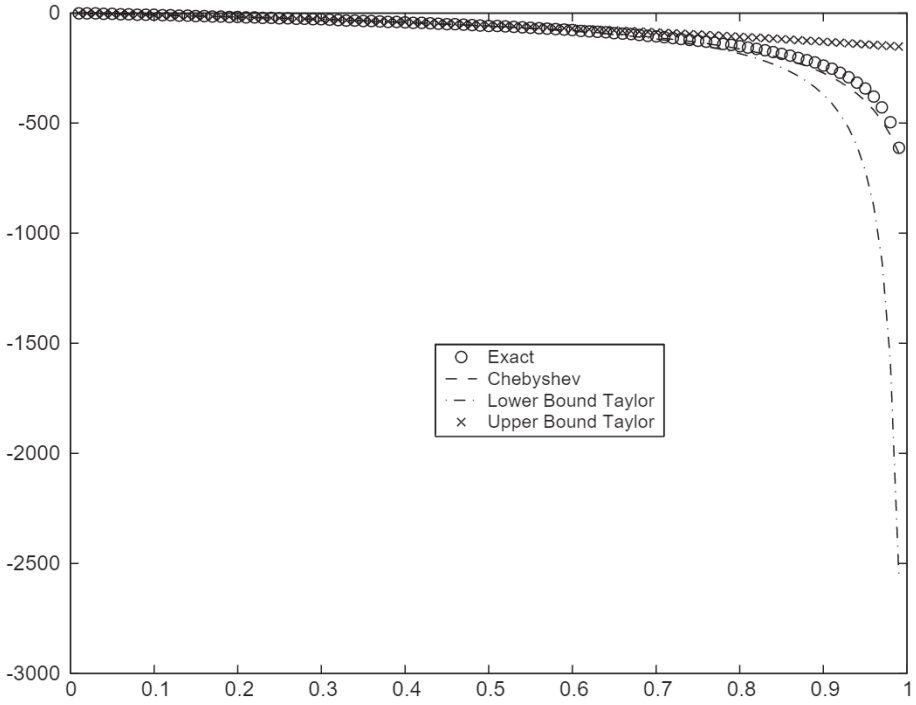

Fig. 3. Differenced exact and approximate log-determinants for US Census tract locations.

If $B(j)_L \leqslant tr(D^j) \leqslant B(j)_U$ for integer $j > q$, $\ln|I-\alpha D|_L \leqslant \ln|I-\alpha D| \leqslant \ln|I-\alpha D|_U$. Hence, appropriate lower and upper bounds for the moments place bounds on the log-determinant.

One can devise several lower bounds for the moments. First, as mentioned by Martin (1993), $tr(D^j) \geqslant 0$ for positive integer j and thus 0 could serve as a lower bound for $tr(D^j)$ when $j > q$.

One can also devise upper bounds on the moments, $tr(D_S^j)$ for $j > q$. By virtue of symmetry and similarity to the row-stochastic D_R, the matrix D_S has real eigenvalues λ_i such that $|\lambda_i| \leqslant 1$ for $i = 1,\ldots,n$. Even powers have all positive eigenvalues since $\lambda_i^{2j} \geqslant 0$ for positive integer j. Suppose the last computed moment q is even and let us claim that $tr(D_S^q) \geqslant tr(D_S^{q+k})$ for integer $k > 0$. We can prove this using induction. For $k = 1$, $tr(D_S^{q+1}) = \sum_{i=1}^n (\lambda_i^q)\lambda_i$ and $\sum_{i=1}^n (\lambda_i^q) \geqslant \sum_{i=1}^n (\lambda_i^q)\lambda_i$, since the maximum value for λ_i is 1 and $\lambda_i^q \geqslant 0$. But $\sum_{i=1}^n (\lambda_i^q) = tr(D_S^q)$ and this shows that $tr(D_S^q) \geqslant tr(D_S^{q+k})$ for $k = 1$. For $k = 2$, $tr(D_S^{q+2}) = \sum_{i=1}^n (\lambda_i^q)\lambda_i^2$ and $\sum_{i=1}^n (\lambda_i^q) \geqslant \sum_{i=1}^n (\lambda_i^q)\lambda_i^2$, since the maximum value for λ_i^2 is 1 and $\lambda_i^q \geqslant 0$. But $\sum_{i=1}^n (\lambda_i^q) = tr(D_S^q)$, and this shows that $tr(D_S^q) \geqslant tr(D_S^{q+k})$ for $k = 2$. For $k = 3,4$ one can define a new even q using $q = q+2$, and use the cases for $k = 1,2$ to show via induction that $tr(D_S^q)$ serves as an upper bound to $tr(D_S^{q+k})$ for integer $k > 0$.

One can construct weight matrices with no decline in the traces of the higher even-powered weight matrices. For example, the very nearest neighbor weight matrix of Pace and Zou (2000) falls into this category. On the other hand, a doubly stochastic matrix with constant off-diagonal elements and zeros on the diagonal has an eigenvalue of 1 and $(n-1)$ eigenvalues equal to $-(n-1)^{-1}$. The traces of the higher order, even-powered weight matrices rapidly converge to the minimum value of 1.

2.5. Efficient computation of the traces

The sparse symmetric structure of D_S means the operation count for computing $tr(D_S^k)$ lies well below $O(n^3)$, especially for small values of k. For example, we know $tr(D_S)=0$ by construction. Let L equal the lower triangle of D_S and note that L has zeros on the main diagonal. Therefore $tr(D_S^2)=tr((L+L')(L+L'))=2\,tr(L'L)$, since $tr(LL)=0$. In addition, $tr(L'L)=\sum\sum(L_{ij})^2$. Let A represent $L'L$ and let B represent LL. Hence,

$$tr(D_S^2)=2\sum\sum(L_{ij})^2$$

and similarly

$$\begin{aligned} tr(D_S^3) &= tr(LLL+L'L'L'+LLL'+L'L'L+2L'LL+2L'LL') \\ &= 6\,tr(L'LL') \\ &= 6\sum\sum A_{ij}L_{ij}. \end{aligned} \tag{6}$$

Eq. (6) relies upon the cyclical redundancy of the trace as well as the invariance of the trace to matrix transposition. Because of these redundancies one could write the above result in several ways. In a similar vein,

$$tr(D_S^4)=2\sum\sum(A_{ij})^2+4\sum(B_{ij})^2+8\sum\sum A_{ij}B_{ij}.$$

Thus, the exact traces use only the matrices L, A, B. Given that L is sparse, A and B will be less sparse due to fill-in. Nevertheless, computing the exact traces seems computationally feasible for large matrices when $k=1,\ldots,4$ and perhaps for higher values of k as well. Computing the exact traces for $k=1,2$ requires little time or storage since $tr(D_S)$ is known a priori and $tr(D_S^2)$ requires only $O(n)$ operations. Other computationally efficient formulae may exist for particular types of weight matrices (e.g., Griffith, 2000).

2.6. Mixed regressive spatially autoregressive model

While a wide variety of spatial models exist, here we consider the mixed regressive spatially autoregressive model (MRSAM) described in Anselin (1988) and Ord (1975). The MRSAM generalizes the spatial autoregression in errors and OLS since appropriate linear restrictions on the MRSAM can be used to produce estimates from both the spatial autoregression in errors model and OLS. In addition, the MRSAM has convenient computational properties. In the MRSAM $y=[U \quad DU \quad \iota]\beta+\alpha DY+\varepsilon$, where y is

the n element vector of observations on the dependent variable, U is the $n \times p$ matrix of observations on the non-constant independent variables, DU is the $n \times p$ matrix of observations on the spatially lagged non-constant independent variables, Dy is the n element vector comprised of the spatially lagged dependent variable, $\imath$ represents a column vector of ones, β is a $2p + 1$ element parameter vector, and α is the scalar spatial autoregressive parameter.

As previously mentioned, the profile likelihood for the case of normally distributed errors is: $L(\alpha) = C + \ln|I - \alpha D| - (n/2)\ln(SSE(\alpha))$. For the MRSAM, $SSE(\alpha) = e'_Y e_Y - 2\alpha e'_{DY} e_Y + \alpha^2 e'_{DY} e_{DY}$ as discussed in Anselin (1988). One can vectorize the log-likelihood by evaluating the profile likelihood for a vector of α values and selecting a value of α corresponding to the largest log-likelihood value, as in Pace and Barry (1997). The use of direct vectorized optimization avoids the overhead and potential fragility of derivative-based optimization algorithms.

Given a vector of log-determinants, it costs little to compute the maximum likelihood estimates for the model and all delete-1 variable sub-models. This permits use of likelihood ratio inference. Note, this means the approximation affects inference and parameter estimation only in its effect upon the location of the maximum of the likelihood function for the various models.

3. Approximation accuracy

This section explores the actual performance of some of the log-determinant approximation techniques discussed in Section 2. In Section 3.1 we compare maximum likelihood estimates of the autoregressive parameter derived from the exact log-determinant and those based on a low-order Chebyshev approximation using a dataset comprised of 57,647 continental US census tracts. In Section 3.2 we calculate maximum likelihood estimates of the autoregressive parameter using a quadratic Chebyshev approximation along with Taylor series bounds for a dataset consisting of 890,091 continental US census blocks.

3.1. Continental US Census tracts

We collected observations on the median price of housing (*Price*), median per capita income (*Income*), median year built, population (*Pop*), the tract's land area (*Area*), as well as the latitude and longitude of the centroid of the tract from the 1990 Census. The variable *Age* equals 1990 less the median year built and was strictly positive. This resulted in 57,647 observations having complete data.

We fitted a mixed regressive spatial autoregressive model to these data where $y = \ln(Price)$ and $U = [\ln(Pop)\ln(Area)\ln(Income)\ln(Age)]$. In specifying the symmetric spatial weight matrix D, we used six weight matrices with different numbers of equally weighted neighbors $(1, 4, 6, 8, 15, 30)$ along with the weight matrix based upon Delaunay triangles.

Table 1
Exact, Chebyshev, and Taylor estimates of the autoregressive parameter for 57,647 US Census tracts

Weight matrix	Max degree polynomial	Exact ML α	Taylor lower bound α	Chebyshev α	Taylor upper bound α
Del	2	0.850	0.770	0.860	0.920
Del	4	0.850	0.830	0.840	0.890
1NN	2	0.610	0.530	0.620	0.680
1NN	4	0.610	0.580	0.610	0.630
4NN	2	0.820	0.740	0.830	0.890
4NN	4	0.820	0.790	0.810	0.850
6NN	2	0.850	0.780	0.870	0.920
6NN	4	0.850	0.830	0.840	0.890
8NN	2	0.870	0.810	0.890	0.940
8NN	4	0.870	0.850	0.860	0.910
15NN	2	0.910	0.850	0.930	0.960
15NN	4	0.910	0.890	0.900	0.940
30NN	2	0.940	0.890	0.960	0.980
30NN	4	0.940	0.920	0.930	0.970

Table 1 presents estimated values of α using both the exact log-determinant and low-order Chebyshev approximation, along with the associated Taylor series bounds.[3] The quadratic Chebyshev approximations resulted in a difference of 0.02 or less in the estimate of α relative to that based on the exact log-determinant. Quartic Chebyshev approximations produced a difference of 0.01 or less compared to using the exact log-determinant. The quadratic Taylor bounds exhibited a maximum width between the upper and lower bound estimates of α equal to 0.15 while the quartic Taylor bounds resulted in a maximum width of 0.06.

To compare the approximate and exact log-determinants, we calculated the exact log-determinants using some of the sparse matrix techniques described in Pace and Barry (1997) on a 1700 Athlon processor. In terms of time, the exact log-determinant for 30 neighbors required 121.14 s, the quartic approximation and bounds took 5.42 s, while the quadratic bounds required only 0.19 s. From these timing results we see that the quadratic Chebyshev approximation to the log-determinant was more than 600 times faster than computing the exact solution.

Table 2 illustrates the use of the log-determinant approximations for specification search. The Chebyshev approximation arrived at the same number of neighbors (15) as the exact solution for both the quadratic and quartic approximations. The upper Taylor series bound also found an optimum at 15 neighbors. However, the Taylor series lower bound found 30 as the optimum number of neighbors for the quadratic and 15 as the optimum number of neighbors for the quartic bounds. One could generalize this specification search by examining weighted nearest neighbor matrices or by examining combinations of weight matrices. As the number of cases rise, the utility of reasonable,

[3] As these estimates come from optimization of the likelihood over a grid of values in 0.01 increments, all the spatial dependence estimates have two digits of accuracy.

Table 2
Specification search based upon exact, Chebyshev, and Taylor techniques (LogLik in thousands)

Weight matrix	Max degree polynomial	Exact Loglik	Taylor lower Loglik	Chebyshev Loglik	Taylor upper Loglik
Del	2	−229.901	−233.338	−229.537	−227.632
Del	4	−229.901	−230.517	−230.146	−229.060
1NN	2	−244.744	−247.689	−244.403	−242.817
1NN	4	−244.744	−245.425	−244.715	−244.331
4NN	2	−231.739	−235.144	−231.178	−229.151
4NN	4	−231.739	−232.377	−231.877	−230.801
6NN	2	−229.532	−232.434	−228.867	−227.120
6NN	4	−229.532	−230.059	−229.763	−228.585
10NN	2	−228.068	−230.277	−227.395	−226.085
10NN	4	−228.068	−228.505	−228.356	−227.221
15NN	2	**−227.522**	−229.255	**−226.909**	**−225.913**
15NN	4	**−227.522**	**−227.905**	**−227.809**	**−226.789**
30NN	2	−227.787	**−228.908**	−227.338	−226.749
30NN	4	−227.787	−228.089	−228.020	−227.265

rapidly computed approximations becomes clear when conducting spatial specification searches.

3.2. Continental US Census blocks

To examine a more computationally challenging data set, we collected observations on White Population, Black Population, Mean Number of Rooms, and Mean Price of housing from individual census blocks in the continental US. This resulted in 890,091 complete data observations. We used a spatial weight matrix based on Delaunay triangles possessing six neighbors per observation on average. We estimated a mixed regressive spatially autoregressive model $y = \ln(\textit{Mean Price})$ and $U = [\ln(\textit{White Pop}) \ln(\textit{White+Black Pop}) \ln(\textit{Mean Rooms})]$ using the quadratic Chebyshev approximation with the corresponding Taylor series bounds, and a least-squares estimate based on the model: $y = [U \quad \iota]\beta + \varepsilon$.

Least-squares results presented in Table 3 indicate that a change in the white population exerts a substantial impact on housing prices, and suggests these differences are statistically significant. Spatial estimates based on the Chebyshev approximation indicate smaller magnitudes for all parameter estimates and signed root deviance statistics.[4] The actual parameter estimate of the effect of white population on housing prices drops in magnitude by over a factor of 3. The spatial model with 8 parameters dramatically outperforms the least-squares model with 4 parameters. RMSE falls from 0.653 for

[4] We employed the signed root deviance statistics (the square root of twice the difference in likelihoods and given the sign of the associated parameter) to avoid scaling issues (Chen and Jennrich, 1996). One can reverse the transformation to yield the original likelihood ratios.

Table 3
Estimates and inference across estimators

Variables	OLS β	Lower bound β	Chebyshev β	Upper bound β
Intercept	8.877	1.779	1.053	0.568
SRD	1020.853	410.833	261.776	145.389
ln(*White Pop %*)	0.184	0.060	0.059	0.058
SRD	239.814	87.342	88.816	88.654
ln(*White + Black Pop*)	0.242	0.062	0.057	0.054
SRD	339.716	150.803	144.665	138.383
ln(*Mean Rooms*)	0.807	0.552	0.547	0.544
SRD	249.437	243.176	250.996	253.207
D ln(*White Pop %*)		−0.011	−0.028	−0.039
SRD		−14.247	−36.464	−51.623
D ln(*White + Black Pop*)		0.050	0.013	−0.012
SRD		81.257	21.289	−21.399
D ln(*Mean Rooms*)		−0.351	−0.411	−0.452
SRD		−117.630	−143.359	−159.494
α		0.770	0.860	0.920
SRD		682.421	725.438	746.118
LogL in Millions	−5.718	−5.200	−5.139	−5.109
RMSE	0.653	0.325	0.312	0.307

OLS to 0.312 for the spatial model estimates based on the quadratic Chebyshev approximation.

Note, the estimates associated with the Taylor series bounds all showed the same sign as the Chebyshev estimates, with the exception of spatially lagged *White + Black Pop* population variable. For this variable, the Taylor upper bound and Chebyshev supported a positive, significant estimate while the Taylor upper bound supported a negative, significant estimate. Thus, the Chebyshev approximation provided information beyond that conveyed by the Taylor bounds. In any event, the quadratic Chebyshev requires the same inputs ($tr(D^2)$) as the Taylor bounds (all O(n)), and so costs little to compute.

To provide another check upon the estimates, we computed the higher order approximation proposed by Barry and Pace (1999).[5] This is a Taylor series approximation using estimated as opposed to exact moments. We examined 50 estimated moments, each estimated with 50 iterations. It required 583.89 s of computation time for the approximation. Using the Monte Carlo approximate log-determinant led to exactly the same estimate of α as the quadratic Chebyshev (0.86), and thus the quadratic Chebyshev and the Monte Carlo log-determinant estimator yielded identical estimates and inferences. In contrast, a simple average of the Taylor bounds led to an estimate of α of 0.82.

Fig. 4 shows the quadratic Chebyshev approximation, the Taylor series lower and upper bounds, and the more precise Monte Carlo log-determinant estimates for the 890,891 by 890,891 Delaunay weight matrix. For small and moderate values of α the

[5] Attempts at exact computation failed due to memory constraints, despite having three gigabytes of RAM.

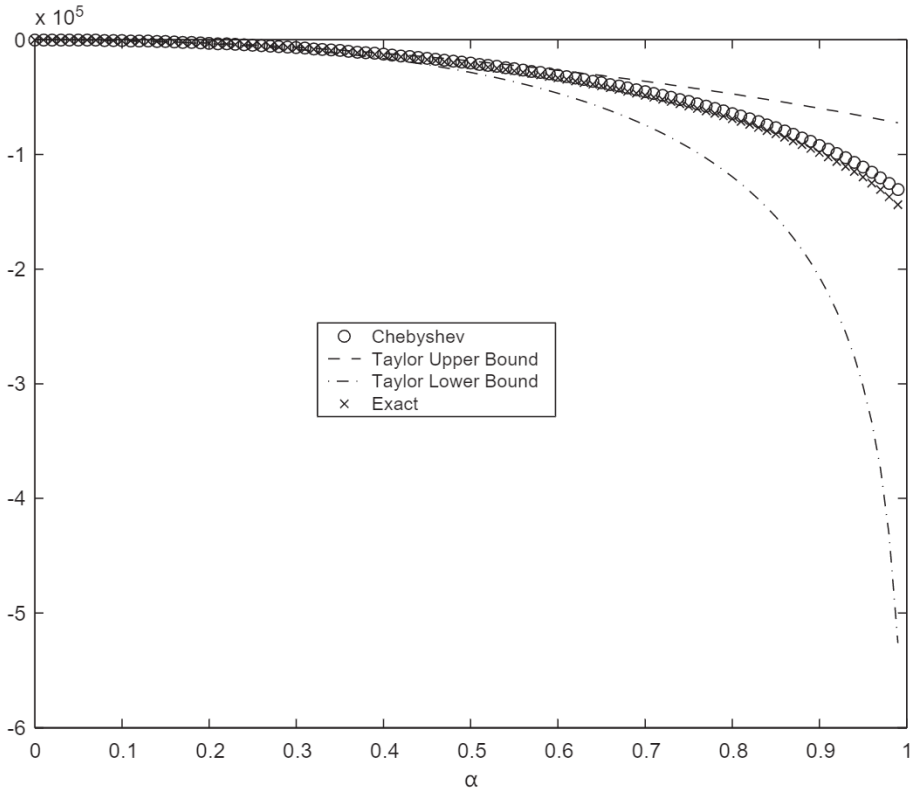

Fig. 4. Approximate log-determinants for US block locations.

curves run together, but they begin to separate at higher levels of α. Even for the larger values of α, the Chebyshev and exact log-determinant curves exhibit a very similar shape while the Taylor series curves diverge. Fig. 5 shows the a differenced version of the same functions.

Note, the Chebyshev approximation was more negative than the exact log-determinant in Fig. 2, but was more positive than the Monte Carlo log-determinant in Fig. 4. Using the average of the upper and lower Taylor bounds as a log-determinant estimator seems appealing based upon Figs. 2 and 3. However, this would result in a substantial downward bias for large values of α based on Figs. 4 and 5, suggesting that such a procedure might not function well in general.

Another point to note is that the log-likelihood value fell within the interval from $-5.200M$ to $-5.109M$ which did not come close to overlapping the exact restricted log-likelihood value of $-5.718M$ from least squares. This indicates that the bounds result in rejection of the restriction, an inference we can draw without computing the exact log-likelihood.

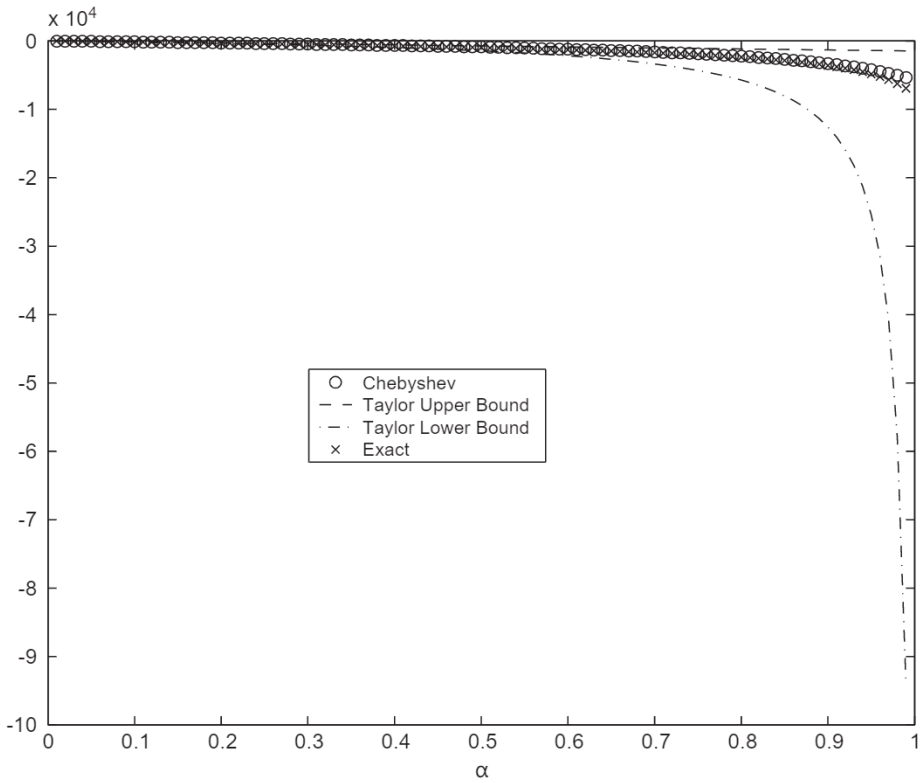

Fig. 5. Differenced approximate log-determinants for US block locations.

It required 318.89 s to create the spatial weight matrix based on Delaunay triangles and only 0.59 s to compute the $tr(D^2)$. Thus, it required under a second to approximate and bound the log-determinant of the $890,091 \times 890,091$ matrix.

4. Conclusion

Recent advances in technology have led to an explosion in the amount of spatial data. For business data alone, the advent of practical geocoding tools have created data sets that far exceed the computational limits of traditional spatial statistical software. However, the approximation techniques proposed here should allow researchers to estimate large data sets as well as conduct more elaborate specification searches without abandoning the elegant likelihood paradigm, of importance to both frequentists and Bayesians.

Taking the Taylor series approximation of Martin (1993) as a point of departure, we use a quadratic Chebyshev approximation to produce estimates of the spatial autoregressive parameter. Computing the exact log-determinant using a sample of 57,647

census tracts with 30 neighbors required 123 s. In contrast, computing the Chebyshev quadratic approximation required only 0.18 s, representing more than a 600-fold speed improvement. This speed improvement becomes progressively better as the sample size increases, because the time required by the exact method rises faster than n (Smirnov and Anselin, 2001 found it rose quadratically in n) while the time required for the Chebyshev quadratic approximation rises linearly with n. The spatial autoregressive parameter estimate differed by only 0.01 or 0.02 from the exact result, a much better performance than expected given the simplicity of the approximation. In another example, computation of the Chebyshev quadratic approximation required less than one second for a sample of 890,091 census blocks, and matched the accuracy of the more elaborate Monte Carlo log-determinant estimator (Barry and Pace, 1999).

While a quadratic approximation to the logarithm function does not perform particularly well for scalars as shown in Fig. 1, the log-determinant approximation fares better. For large matrices, the log-determinant is a sum of univariate approximations, with arguments equal to the eigenvalues of the matrix. Since the Chebyshev approximation has errors of opposing signs, the sum of these errors behaves better than one might suspect given Fig. 1 as illustrated by Figs. 2 and 4. So the Chebyshev approximation appears to perform well for larger matrices, its intended application.

Many approximations provide little information on their accuracy. Using bounds on the traces of higher order matrix powers enabled a Taylor series to bound the accuracy of the Chebyshev approximation. As a result, the user obtains a lower bound on the log-determinant, an approximate log-determinant, and an upper bound on the log-determinant. Using likelihood dominance arguments, if a restricted log-likelihood is in the interval of say $[-2000, -1500]$ compared to an unrestricted log-likelihood in the interval $[-1200, -800]$, one can reject a reasonable number of restrictions without knowing the precise value of the log-likelihood.

While a tremendous amount of data exhibits substantial spatial dependence, users of data in many fields still estimate models using aspatial techniques. This becomes especially pertinent in the area of spatial data mining where vast quantities of data often lead users to choose ad hoc, non-traditional estimators for their applications. The simple Chebyshev approximation to the log-determinant term in the spatial likelihood function should allow users to employ standard maximum likelihood or Bayesian methods for their problems, extending the benefits of spatial estimation to a larger class of applications.

Acknowledgements

We would like to gratefully acknowledge the research support received from the National Science Foundation (BCS-0136193 and BCS-0136229).

References

Anselin, L., 1988. Spatial Econometrics: Methods and Models. Kluwer Academic Publishers, Dordrecht.

Barry, R., Pace, R.K., 1999. A Monte Carlo estimator of the log determinant of large sparse matrices. Linear Algebra Appl. 289, 41–54.

Chen, J.-S., Jennrich, R., 1996. The signed root deviance profile and confidence intervals in maximum likelihood analysis. J. Amer. Statist. Assoc. 91, 993–998.

Francica, J., 2000. Large spatial databases. Bus. Geogr. 8, 18–23.

Golub, G., Van Loan, C., 1996. Matrix Computations, 3rd Edition. John Hopkins Press, Baltimore.

Griffith, D., 2000. Eigenfunction properties and approximations of selected incidence matrices employed in spatial analyses. Linear Algebra Appl. 321, 95–112.

Griffith, D., Sone, A., 1995. Trade-offs associated with normalizing constant computational simplifications for estimating spatial statistical models. J. Statist. Comput. Simulation 51, 165–183.

Kelejian, H., Prucha, I.R., 1998. A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. J. Real Estate Finance Econom. 17 (1), 99–121.

LeSage, J.P., 1997. Bayesian estimation of spatial autoregressive models. Internat. Regional Sci. Rev. 20, 113–129.

LeSage, J.P., 2000. Bayesian estimation of limited dependent variable spatial autoregressive models. Geogr. Anal. 32, 19–35.

Marcus, M., Minc, H., 1992. A Survey of Matrix Theory and Matrix Inequalities. Dover, New York.

Martin, R.J., 1993. Approximations to the determinant term in Gaussian maximum likelihood estimation of some spatial models. Comm. Statist. Theory Methods 22, 189–205.

Muller, J.-M., 1997. Elementary Functions. Birkhäuser, Boston.

Ord, J.K., 1975. Estimation methods for models of spatial interaction. J. Am. Statist. Assoc. 70, 120–126.

Pace, R.K., Barry, R., 1997. Quick computation of regressions with a spatially autoregressive dependent variable. Geogr. Anal. 29, 232–247.

Pace, R.K., Zou, D., 2000. Closed-form maximum likelihood estimates of nearest neighbor spatial dependence. Geogr. Anal. 32, 2000.

Press, W., et al., 1996. Numerical Recipes in Fortran 77. Cambridge University Press, New York.

Smirnov, O., Anselin, L., 2001. Fast maximum likelihood estimation of very large spatial autoregressive models: a characteristic polynomial approach. Comput. Statist. Data Anal. 35, 301–319.

Toh, K.-C., Trefethen, L.N., 1998. The Chebyshev polynomials of a matrix. SIAM J. Matrix Anal. Appl. 20, 400–419.

Whittle, P., 1954. On stationary processes in the plane. Biometrika 41, 434–439.