ELSEVIER

Statistics & Probability Letters 33 (1997) 291–297

STATISTICS & PROBABILITY LETTERS

Sparse spatial autoregressions

R. Kelley Pace[a], Ronald Barry[b,*]

[a] *Department of Finance, School of Management, University of Alaska, Fairbanks, AK 99775-6080, USA*
[b] *Department of Mathematical Sciences, University of Alaska, Fairbanks, AK 99775-6660, USA*

Received March 1996; revised May 1996

Abstract

Given local spatial error dependence, one can construct sparse spatial weight matrices. As an illustration of the power of such sparse structures, we computed a simultaneous autoregression using 20 640 observations in under 19 min despite needing to compute a 20 640 by 20 640 determinant 10 times.

Keywords: Spatial autoregression; SAR; Sparse matrices

1. Introduction

Regression is perhaps the most often used technique in statistics. When applied to spatially distributed observations, however, much predictive power can be lost by ignoring the presence of spatial autocorrelation. Moreover, ignoring spatial autocorrelation leads a serious violation of the assumptions underlying ordinary least squares regression which can result in erroneous statistical inference. Fortunately, a variety of spatial estimators can adjust for this problem. Example include simultaneous autoregressions (SAR), conditional autoregressions (CAR), and kriging (see Cressie, 1993).

Unfortunately, all of these techniques involve examining the explicit relation between an observation and all other observations. If n observations exist, this leads to n^2 potential relations. Hence, as n becomes large, computing spatial estimators can become quite expensive as these usually require computing the determinant or inverse of an $n \times n$ matrix which requires order of n^3 operations ($O(n^3)$). This conflicts with the increasing prevalence of large data sets involving thousands of observations. Clearly, standard spatial statistical methods can become impractical for many realistic applications.

Fortunately, spatial autocorrelation usually declines with distance. Truncating the influence of observations past a certain distance could greatly reduce the number of relations (non-zero elements) needed to estimate the spatial regression. Mathematically, sparse matrices can represent this situation. Sparse matrix

* Corresponding author. Tel.: (907)-474-7226; fax: (907)-474-5394; e-mail: FFRPB@aurora.alaska.edu.

PII S0167-7152(96)00140-X

techniques store only the non-zero elements of a matrix and also avoid performing unnecessary computations on the zero elements. This both reduces storage space and accelerates execution time.[1]

To illustrate the dramatic improvements possible, we computed a simultaneous autoregression (SAR) using 20 640 observations. Each SAR likelihood function evaluation requires computing the determinant of the 20 640 by 20 640 matrix. The sparsity of the problem (0.019% non-zero elements) makes it possible to compute ten evaluations of the likelihood in under 19 min. Moreover, a dense solution to the problem would have required storing around 3.4 GB of data while the sparse solution to the problem used just over one megabyte of memory.

To place these results into perspective, Li (1995) used a IBM RS6000 Model 550 and a CM5 parallel processing supercomputer to compute a 2500 observation SAR. The CM5 had 32 processors each with 32 MB of local memory and four vector units. For a 2500 by 2500 spatial weight matrix the RS6000 required 8515.07 s while the CM5 required 48 s. Since computing determinants requires $O(n^3)$ operations, the differences in size would require a factor of $(20\,640/2500)^3$ more operations, assuming no additional bottlenecks. Given this size adjustment, the extrapolated computation times would go to over 55 d for the RS6000 and 7.5 h for the CM5 on the 20 640 by 20 640 problem. Hence, personal computers with sparse technology can exceed supercomputer performance for dense technology.

The tremendous gains in computing speed do not come at the expense of statistical performance. For the simple model of California housing prices examined, the SAR manages to achieve a median absolute error of 0.1084 while OLS produced a median absolute error of 0.2101, almost twice as large.

Section 2 discusses the spatial autoregressive error estimator employed, Section 3 provides details on an improved algorithm for computing spatial estimators, Section 4 estimates the resulting spatial error autoregression, while Section 5 concludes with the key results.

2. A spatial autoregressive error estimator

Section 2.1 describes the likelihood function for a spatial autoregressive error process. This likelihood function depends critically upon D, the spatial weighting matrix. Hence, Section 2.2 describes D in more detail.

2.1. The spatial autoregressive error likelihood function

When errors exhibit spatial autocorrelation, the simultaneous autoregression (SAR) estimator corrects the usual prediction of the dependent variable, $Y = X\beta + \varepsilon$, by a weighted average of the 'deviations' $Y - X\beta$ on nearby observations as in (1),

$$Y = X\beta + \alpha D(Y - X\beta) + \varepsilon \tag{1}$$

where D represents an $n \times n$ weighting matrix with 0's on the diagonal (the observation cannot predict itself) and non-negative off-diagonals. To maintain the interpretation of a weighted average, the rows of D sum to 1 as implied by (2) below. Such weighting matrices are said to be row-standardized (Haining, 1990, p. 82). A non-zero entry in the jth column of the ith row indicates that the jth observation will be used to adjust the prediction of the ith observation ($i \neq j$). After correcting for these interactions, the SAR models assume the

[1] See George and Liu (1981), Golub and Van Loan (1989), Press *et al.* (1988), and Saad (1996) for general discussions of sparse matrix techniques.

residuals, ε, are independently and normally distributed. These assumptions appear in (2).

$$\underset{(n\times n)}{D}\ \underset{(n\times 1)}{[1]} = \underset{(n\times 1)}{[1]} \tag{2a}$$

$$\operatorname{diag}(D) = \underset{(n\times 1)}{[0]} \tag{2b}$$

$$0 \leqslant \alpha < 1 \tag{2c}$$

$$\varepsilon \sim \mathrm{N}(0, \sigma^2 I) \tag{2d}$$

The simultaneous autoregression (SAR) has the following log-likelihood function:

$$L|(\alpha, \beta, \sigma^2) = \tfrac{1}{2}\ln|B| - \tfrac{1}{2}[n\ln(2\pi\sigma^2) + \sigma^{-2}(Y - X\beta)'B(Y - X\beta)], \tag{3}$$

where B equals $(I - \alpha D)'(I - \alpha D)$.[2] To ensure the sum-of-squared errors, $(Y - X\beta)'B(Y - X\beta)$, is strictly positive, B must be positive-definite. Given the definition of D, for $1 > \alpha \geqslant 0$, $1 \geqslant |B| > 0$, and hence $0 \geqslant \ln|B| > -\infty$. The maximum likelihood method efficiently estimates the model asymptotically (given the assumptions hold).

Assuming the existence of the ML estimate, one could predict Y via (4).

$$\breve{Y} = X\breve{\beta} + \breve{\alpha}D(Y - X\breve{\beta}). \tag{4}$$

Furthermore, (4) leads to the estimated errors in (5).

$$\breve{e} = Y - \breve{Y} = Y - X\breve{\beta} - \breve{\alpha}D(Y - X\breve{\beta}) = (I - \breve{\alpha}D)(Y - X\breve{\beta}) \tag{5}$$

2.2. *Specification of the spatial weight matrix*

The weight given to the census block groups for differencing depends upon their proximity as measured by the latitude and longitude for each observation relative to all other observations. Let d_{ij} represent the Euclidean distance between every pair of observations i and j and let $d_{\max i}$ represent the distance between the ith observation and its mth nearest neighbor. Moreover, let $w_{ij} = 1$ if $d_{ij} \leqslant d_{\max i}$ and zero otherwise as stated by (6).

$$d_{ij} \leqslant d_{\max i} \leftrightarrow w_{ij} = 1\,. \tag{6}$$

Naturally, this yields a weight of 1 for the census block group itself ($d_{ij} = 0$) and 0 for each observation j more than $d_{\max i}$ distance from observation i. Subsequently, in (7) we normalized the initial weights so that $\sum_{j=1, i\neq j}^{n} D_{ij} = 1$ thus making it into a standardized weight matrix.

$$D_{ij} = \frac{w_{ij}}{\sum_{j=1, i\neq j}^{n} w_{ij}} \tag{7}$$

In addition, we set $D_{ii} = 0$, as assumed in (2), to prevent each observation from predicting itself.

For the third observation, D might appear as,

$$D_{3,1:15810} = [0, 0.25, 0, 0, 0.25, 0, 0, 0, 0.25, 0, 0.25, 0, \ldots, 0]\,.$$

Note, the third entry of $D_{3,1:15810}$ equals 0 while the row sums to 1.

---

[2] For example, see Cressie (1993, p. 465). See Pace and Gilley (forthcoming) for an alternative application.

3. Computation of spatially autocorrelated error regressions

Examination of (3) shows the main barrier to speedy computation of the estimates lies in the $n \times n$ nature of the spatial weighting matrix, D. In particular, computing $|I - \alpha D|$, a determinant of an $n \times n$ matrix, requires substantial time for large n. Standard determinant computations use $O(n^3)$ operations while multiplication of a $n \times n$ matrix and an $n \times p$ matrix uses $O(n^2 p)$.[3] Thus, the storage requirements, temporary or permanent, rise with the square of n while the operation count rises with the cube of n. A problem with many observations would quickly exhaust storage space or require impractical amounts of computing time.

Fortunately, one can avoid such Herculean computational tasks. The spatial weighting matrix D contains mainly 0's and hence is sparse. The spatial weight matrix described in Section 2.2 uses the m nearest neighbors out of a possible n neighbors for each of the n observations. Thus, mn non-zero entries exist out of a possible n^2 elements. A common measure of matrix sparsity examines the number of non-zero elements of a matrix relative to the total number of elements. Hence, the m nearest neighbor weighting matrix D has proportionally m/n non-zero elements (i.e., 4/20 640). D is quite sparse for this problem and becomes progressively more sparse for large empirical applications.

This level of sparseness totally alters the number of operation counts needed for the computation of determinants and matrix multiplication. The operation count depends more upon the number of non-zero elements than the total number of elements. Moreover, the sparse matrix procedures store only the non-zero elements and some form of index. Hence, sparse matrix methods can make it possible to handle very large applications.

Sparse methods fall into two basic categories, direct and iterative. Statisticians commonly use iterative methods such as the conjugate gradient in the optimization of non-linear problems. However, the conjugate gradient, as well as a host of other techniques such as the Jacobi, successive overrelaxation, Krylov subspace methods, and so forth, also have applications in linear systems. A very sparse matrix makes it extremely fast to compute each iteration. Like any minimization (maximization) problem, the algorithms need positive (negative) definite matrices to ensure a global solution. Moreover, the more pronounced the positive (negative) definiteness, the faster convergence usually proceeds. See Saad (1996) for more details.

Direct methods take advantage of blocks of zeros in the gaussian elimination process. Hence, direct methods usually prefer contiguous groups of zeros in the matrix. Specific patterns such as bandedness and bordered matrices have dedicated methods for their solution (Press et al., 1988, pp. 72–74). More general patterns require the use of factorization methods such as the reverse Cuthill–McKee, minimum degree (followed here), and nested dissection, to reorder the matrix for fast gaussian elimination (George and Liu, 1981).

Given the close connection of the determinant with gaussian elimination, computing spatial autoregressions seems more suited to direct methods. A potential problem arises, however, as the goal of reordering the matrix to maximize computation speed could conflict with the ordering (pivoting) of the matrix to maximize numerical stability. Fortunately, the structure of the problem leads to either diagonal dominance or symmetry. Both situations have very favourable error properties and do not require pivoting. Examination of $(I - \alpha D)$ shows for $\alpha < 1$ the sum of the off-diagonal elements in each row is less than the diagonal (1). Hence, for $(I - \alpha D)'$ the sum of the off-diagonal elements in each colomn is less than the diagonal (1) and $(I - \alpha D)'$ is diagonal dominant. Since $|(I - \alpha D)| = |(I - \alpha D)'|$, computing $|(I - \alpha D)'|$ sidesteps the error issues which arise from reordering (Golub and Van Loan, 1989, p. 119–120). Also, we could compute $|(I - \alpha D)'(I - \alpha D)| = |(I - \alpha D)'||(I - \alpha D)| = 2|(I - \alpha D)|$. The symmetry and positive definiteness of $|(I - \alpha D)'(I - \alpha D)|$ allows the use of the Cholesky decomposition which also does not need pivoting to achieve numerical stability (George and Liu, 1981, p. 9).

[3] See Golub and Van Loan (1989, p. 99).

We used the MATLAB programming language running on a 133 MHz Pentium computer to generate the estimates. The sparse matrix formulation required only 1130 s to evaluate 10 iterations of the likelihood function. Each iteration involved computing the determinant of a 20 640 by 20 640 matrix as well as various multiplications. In addition, the storage of the matrix took somewhat over 1 MB whereas the full matrix would have required 3.4 GB of memory.

These speed increases coupled with decreases in storage requirements make the estimation of large spatial problems practical. For medium and smaller problems it allows users to jointly model other phenomenon of interest such as specification or simultaneity.

4. Maximum likelihood sample estimation of a spatial autoregression

This section illustrates the spatial autoregression estimator from Section 2 using the 1990 census data. Section A discusses the model, Section 2 presents the data, and Section 3 presents the actual estimation results.

4.1. Model

We fitted the following model:

$$\begin{aligned}\ln(\text{MEDIAN VALUE}) = {} & \text{INTERCEPT} + \beta_2\,\text{MEDIAN INCOME} + \beta_3\,\text{MEDIAN INCOME}^2 + \beta_4\,\text{MEDIAN INCOME}^3 \\ & + \beta_5 \ln(\text{MEDIAN(AGE)}) + \beta_6 \ln(\text{TOTAL ROOMS/POPULATION}) \\ & + \beta_7 \ln(\text{BEDROOMS/POPULATION}) + \beta_8 \ln(\text{POPULATION/HOUSEHOLDS}) \\ & + \beta_9 \ln(\text{HOUSEHOLDS}) \end{aligned} \tag{8}$$

4.2. Data

We collected information on the variables in (8) using all the block groups in California from the 1990 Census. In this sample a block group on average includes 1425.5 individuals living in a geographically compact area. Naturally, the geographical area included varies inversely with the population density. We computed distances among the centroids of each block group as measured in latitude and longitude. We excluded all the block groups reporting zero entries for the independent and dependent variables. The final data contained 20 640 observations on 9 characteristics.

4.3. Maximum likelihood sample estimation

Table 1 contains the sample estimates from using OLS and the SAR maximum likelihood estimators. To emphasize the sparsity of the problem a priori we picked four nearest neighbors to receive positive weights in the spatial weigth matrix D. Based upon a numerical search, the SAR maximum likelihood estimate of α was 0.8536.

Note the treatment by the two estimators of the AGE variable. OLS produces a positive and significant estimate of the AGE variable with a t statistic of 33.6133 while the maximum likelihood SAR produces a negative, significant estimate with a t statistic of -11.0942.[4]

[4] See Pace and Gilley (1993) and Gilley and Pace (1995) for a discussion of priors in hedonic pricing models.

Table 1
OLS and SAR estimates for median housing prices across 20 640 California census block groups

	B_{ols}	t_{ols}	B_{sar}	t_{sar}
INTERCEPT	11.4939	275.7518	11.6637	402.5925
MEDIAN INCOME	0.4790	45.7768	0.0349	4.7104
MEDIAN INCOME^2	−0.0166	−9.4841	0.0100	8.4280
MEDIAN INCOME^3	−0.0002	−1.9157	−0.0007	−12.2444
ln(MEDIAN AGE)	0.1570	33.6123	−0.0421	−11.0942
ln(TOTAL ROOMS/POPULATION)	−0.8582	−56.1280	0.3098	24.5768
ln(BEDROOMS/POPULATION)	0.8043	38.0685	−0.1926	−11.8049
ln(POPULATION/HOUSEHOLDS)	−0.4077	−20.8762	−0.0342	−2.3582
ln(HOUSEHOLDS)	0.0477	13.0792	0.0034	1.5569
α			0.8536	
R^2	0.6078		0.8594	
Median $\|e\|$	0.2101		0.1084	
Execution time			1130 s	
Number of likelihood evaluations			10	

Even when both estimators agree in terms of the direction of an effect, they may differ in their implications concerning the functional form governing the effect. For example, OLS arrives at a linear estimate of the effects of MEDIAN INCOME upon ln(PRICE) with negative quadratic and cubic effects. In contrast, the SAR estimate shows a positive linear and quadratic effect of MEDIAN INCOME upon ln(PRICE) with a negative cubic effect.

For many applications prediction error constitutes the main concern. In this respect the SAR estimator greatly outperforms OLS. For example, the OLS sample R^2 was 0.6078 while for the SAR estimate the sample R^2 was 0.8536, a dramatically better goodness of fit. Equally dramatic, the median absolute errors under OLS of 0.2101 fall by 48.4% to 0.1084 under the SAR estimator.

For other applications statistical inference constitutes the main concern. The presence of such pervasive spatial autocorrelation completely invalidates the use of the usual iid based OLS p-values. Indeed, the dramatic change in many of the t-values between the two estimators emphasizes how conditional these are upon the assumed correlation structure.

5. Conclusion

Accurate prediction and correct inference in a spatial setting must use the information contained in the errors on nearby observations. The canonical estimators available, however, require order of n^3 computations. Given many applications must deal with thousands of observations, this presents a substantial barrier to the adoption of the standard spatial techniques. However, as a stylized fact, only the relations among nearby observations matter greatly which means one can employ sparse matrix techniques to achieve a great savings of both storage space and execution time.

As an illustration of the power of these techniques, the 20 640 observations employed herein would have required the use of a matrix taking 3.4 GB using normal dense methods. The sparse matrix formulation of the problem droped the storage space to just above 1 MB. Finding the spatial maximum likelihood estimate required 10 likelihood evaluations each needing the computation of the determinant of a $20\,640 \times 20\,640$ matrix. This took under 19 min on a 133 MHz Pentium computer.

As expected, incorporating the spatial information greatly reduced prediction errors. For the simple model employed, the spatial estimator displayed a median absolute error almost one-half less than OLS. Moreover, the dramatic shift in t-values between the two estimators emphasized how conditional these are upon the assumed correlation structure.

The great gains produced by sparsity pose the question of what statistical factors lead to sparsity. First, if the regression function performs perfectly, it will leave only white noise and I would be the optimal spatial weight matrix. At the other extreme, a poor regression function might produce slowly decaying correlated errors and greatly reduce the sparsity of the spatial weight matrix. Often autocorrelated errors arise becuase of omitted variables. For example, pollution from a factory might decline with the inverse of the squared distance. Applying a model without a pollution independent variable might result in correlated errors decaying with the inverse of the squared distance. Second, the span and density of the data influence sparsity. An optimal spatial weighting matrix might include more neighboring observations in predicting Hong Kong apartment rents than in predicting suburban Houston apartment rents. On a national scale the neighboring observations for any given apartment would occupy a small relative bandwidth relative to the equivalent problem over one census tract. Sparsity of various forms arise in many physical and electronic systems (Rice, 1981, pp. 25–28).

The advent of geographical information systems which provide data sets of ever-increasing size, the superior prediction as well as inference from spatial estimators using this information, and the low-cost computation of spatial estimators using sparse matrices have the potential to greatly increase the appeal of spatial estimators.

Acknowledgements

We would like to gratefully acknowledge the research support we have received from the University of Alaska. We would also like to thank Otis Gilley, an anonymous referee, and the editor for their thoughtful comments as well as the seminar participants at the University of Connecticut at Storrs and Southern Illinois University at Edwardsville.

References

Cressie, N.A.C. (1993), *Statistics for Spatial Data* (Wiley, New York, 2nd ed.).

George, A. and J. Liu (1981), *Computer Solution of Large Sparse Positive Definite Systems* (Prentice-Hall, Englewood Cliffs, NJ).

Gilley, O.W. and R.K. Pace (1995), Improving hedonic estimation with an inequality restricted estimator, *Rev. Econom. Statist.* 77, 609–621.

Golub, G.H. and C.F. Van Loan (1989), *Matrix Computations* (John Hopkins, Baltimore, 2nd ed.).

Haining, R. (1990), *Spatial Data Analysis in the Social and Environmental Sciences* (Cambridge University Press, Cambridge).

Li, B. (1995), Implementing Spatial Statistics on Parallel Computers, in: S. Arlinghaus, ed. *Practical Handbook of Spatial Statistics* (CRC Press, Boca Raton), pp. 107–148.

Pace, R.K. and O.W. Gilley (1993), Improving prediction and assessing specification quality in non-linear statistical valuation models, *J. Business Econom. Statist.* **11**, 301–310.

Pace, R.K. and O.W. Gilley, Using the spatial configuration of the data to improve estimation, *J. Real Estate Finance Econom.*, forthcoming.

Press *et al.* (1988), *Numerical Recipes in C* (Cambridge University Press, Cambridge).

Rice, J. (1981), *Matrix Computations and Mathematical Software* (McGraw-Hill, New York).

Saad, Y. (1996), *Iterative Methods for Sparse Linear Systems* (PWS Publishing, Boston).