Lecture 6: Matrix Exponential Spatial models

James P. LeSage
University of Toledo
Department of Economics
Toledo, OH 43606
jlesage@spatial-econometrics.com

March 2004

1 Introduction

Estimation of traditional spatial autoregressive (SAR) models requires non-linear optimization for estimation and inference. The conventional spatial autoregressive approach introduces additional theoretical complexity relative to non-spatial autoregressive models and is difficult to implement in large samples. We advocate use of a matrix exponential spatial specification (MESS) of dependence that replaces the conventional geometric decay of influence over space with an exponential pattern of decay. We show that this results in theoretical simplicity as well as improved numerical performance relative to the conventional spatial autoregression.

Chiu, Leonard, and Tsui (1996) proposed the use of the matrix exponential for covariance matrix modelling and discussed several of its advantages. One advantage is that the matrix exponential always leads to positive definite covariance matrices, eliminating the need to restrict the parameter space or test for positive definiteness during optimization. A second advantage is that inversion of the matrix exponential takes a simple mathematical form that is easy to implement in applied practice. Finally, use of the matrix exponential spatial specification leads to a log-likelihood where a troublesome term involving the log-determinant of an nxn covariance matrix vanishes. Collectively, these aspects of the matrix exponential spatial specification greatly simplify maximum likelihood as well as Bayesian estimation and inference. Specifically, we are able to provide a closed-form solution for maximum likelihood estimates, and produce Bayesian estimates using univariate integration over a scalar polynomial expression. In addition, we show how MESS can be used for model diagnostics and comparison of models based on different spatial weight structures or sets of explanatory variables. We demonstrate these procedures using a number of data sets that vary in size and area of application.

Section 2 presents the MESS model, associated likelihood, and means of finding closed-form maximum likelihood estimates. Bayesian estimation of the model is discussed in section 3, and Bayesian model comparison is taken up in section 4. Section 5 provides applied illustrations of the various techniques using a number of spatial data sets.

2 The MESS model and maximum likelihood solutions

Section 2.1 sets forth the matrix exponential spatial specification, and section 2.2 compares the data generating processes associated with a conventional SAR model to the MESS data generating process. Section 2.3 presents the log-likelihood and simplifications that follow from characteristics of the matrix exponential, while section 2.4 provides a closed-form solution for MESS model parameter estimates. Section 2.5 discusses estimates of dispersion for the parameters and section 2.6 gives means of computing model diagnostics for the MESS.

2.1 The Matrix exponential spatial specification

We begin by assuming the dependent variable vector y exhibits spatial dependence such that observation i may depend on neighboring observations, $j \in \aleph$, where $\aleph$ denotes a set of neighboring observations. We model this spatial dependence using a linear transformation Sy of the dependent variable y, as in (1) or (2).

$$Sy = X\beta + \varepsilon \tag{1}$$
$$y = S^{-1}X\beta + S^{-1}\varepsilon \tag{2}$$

The vector y contains the n observations on the dependent variable, X represents the nxk matrix of observations on the independent variables, and the n-element vector ε is distributed $N(0, \sigma^2 I_n)$. The nxn matrix S is positive definite. One obtains the traditional SAR model when $S = (I - \rho D)$, where D represents an nxn non-negative spatial weight matrix. Elements of D_{ij} are set to positive values for observations $j \in \aleph$, and by convention, $D_{ii} = 0$. Often, D is row-stochastic such that $D\iota = \iota$, where ι denotes a vector of ones. Row-stochastic spatial weight matrices, or multidimensional linear filters, have a long history of application in spatial statistics (e.g., Ord (1975)). The row-stochastic weight matrix has favorable statistical, numeric, and interpretive properties. For example, the product of a row-stochastic weight matrix D and a random variable vector v produces a vector of spatial local averages, Dv.

Instead of the spatial autoregressive definition of $S = (I - \rho D)$, we propose using,

$$S = e^{\alpha D} = \sum_{t=0}^{\infty} \frac{\alpha^t D^t}{t!} \tag{3}$$

with α denoting a scalar parameter. In other words, we advocate using a matrix exponential transformation to model spatial dependence. While this appears quite different from the conventional spatial autoregressive model, we will show that the similarities outweigh the differences. Although the matrix exponential transformation behaves in a similar fashion as the spatial autoregressive specification, it has a number of advantages. In fact, Chiu, Leonard, and Tsui (1996) proposed the use of the matrix exponential for covariance matrix modelling and discussed several of its salient properties, some of which are:

Property 1: S is positive definite,

Property 2: $S^{-1} = (e^{\alpha D})^{-1} = e^{-\alpha D}$,

Property 3: $|e^{\alpha D}| = e^{\text{trace}(\alpha D)}$.

These properties lead to a number of practical advantages. Property 1 indicates that the matrix exponential leads to positive definite covariance matrices, and thus avoids the need to restrict the parameter space, or to carry out tests for positive definiteness during parameter estimation. Property 2 leads to simple mathematical inversion of the matrix exponential and correspondingly simple numerical inversion procedures, which benefit both theoretical and applied work. It will be shown in 2.3, that Property 3 eliminates a troublesome log-determinant involving an nxn covariance matrix from the log-likelihood of the MESS model.

2.2 A comparison with spatial autoregressive models

The data generating process (DGP) for the traditional SAR model can be expressed as in (4), having expectation $E(y_s)$ shown in (5).

$$y_s = (I_n - \rho D)^{-1} X\beta + (I_n - \rho D)^{-1}\varepsilon \tag{4}$$

$$E(y_s) = (I_n - \rho D)^{-1} X\beta = \sum_{i=0}^{\infty} \rho^i D^i X\beta \tag{5}$$

Theoretical economic models have been used to justify this type of DGP for cases involving spatial spillovers, spatial competition and latent unobservable variables with spatial dependence (Brueckner, 2002), and for production processes in a spatial context (López-Bazo, et al., 1999). We provide one justification of the MESS model DGP by demonstrating a close correspondence between the expectation $E(y_m)$ from the MESS model and $E(y_s)$ of the conventional spatial autoregressive model. Using Property 2 above, $S^{-1} = e^{-\alpha D}$, we can express the data generating process for the MESS model as in (6), with expectation $E(y_m)$ in (7).

$$y_m = S^{-1} X\beta + S^{-1}\varepsilon \tag{6}$$

$$E(y_m) = S^{-1} X\beta = \sum_{i=0}^{\infty} \frac{(-\alpha)^i D^i}{i!} X\beta \tag{7}$$

One approach to specifying the dependence structure in the spatial connectivity matrix D is to rely on non-zero weights assigned to some number m of nearest neighbor observations. In this situation, the spatial weight matrix D operates to produce dependence of individual observations on the m nearest neighbors. Powers of this weight matrix reflect neighbors to these m nearest neighbors, so that the ith row of D^2 contains positive values for neighbors to the m nearest neighbors to observation i. Similar relations hold for higher powers of D that identify higher-order neighbors.

The spatial weight matrix D in conventional specifications is often row-standardized to have row-sums of unity. Since products of row-stochastic matrices are row-stochastic, we have by definition that $D\iota = \iota$, $D(D\iota) = \iota$, and so on. The same holds true for any power of S, since the powers are simply linear combinations of the powers of D, all of which are proportional to a row-stochastic matrix. Thus, if D is row-stochastic, $e^{\alpha D}$ will be proportional to a row-stochastic matrix. In fact, the row sums equal e^{α}.

To illustrate the close connection between the MESS and SAR DGP, we used a number of spatial data samples based on $n = 49$, 3,107, 22,210, and 57,188 observations and a spatial weight matrix based on the five nearest neighbors to examine the correspondence between the vectors $E(y_s)$ and $E(y_m)$.[1] The DGP for the SAR model was used to generate $E(y_s)$ over a 0.01 grid of values for ρ between -0.99 and 0.99. Using the DGP for the MESS model

[1] The 49 observation data set was constructed using latitude-longitude centroids from Anselin (1988) on neighborhoods from Columbus, Ohio, the 3,107 observation sample was constructed from US counties, the 22,210 observation sample used US agricultural zip-code areas, and the 57,188 observation sample was based on US census tracts.

to produce $E(y_m)$, we solved for the value of α that minimized the sum of squared errors between $E(y_m)$ and $E(y_s)$. Values of α between -2.75 and 1.0 were capable of producing $E(y_m)$ such that the $R-$squared between $E(y_m)$ and $E(y_s)$ was 0.99 in the interval from $\rho = -0.99$ to $\rho = 0.8$, and above 0.98 for the remaining values of ρ. This correspondence is not surprising given that we are replacing the SAR model geometric pattern of decay with a flexible pattern of exponential decay.

The relation between α and ρ values suggests a correspondence, $\rho = 1 - \exp(\alpha)$, between the traditional spatial dependence parameter ρ and α. A plot of this correspondence is shown in Figure 1, where results based on the four spatial samples used in the experiment are presented. The correspondence is exact for $\rho = 0$, allowing us to interpret $\alpha = 0$ as indicative of no spatial dependence, and we note that negative values for α correspond to positive spatial dependence ($\rho > 0$), with positive values indicating negative dependence ($\rho < 0$). This should serve as a useful rule-of-thumb for practitioners, allowing translation of α estimates of the spatial dependence parameter to the traditional spatial autocorrelation scale.

2.3 MESS log-likelihood

The log-likelihood for the MESS model is in (8), and a profile log-likelihood where β and σ have been concentrated out of the likelihood is in (9).

$$
\begin{aligned}
\ln L(\beta, \sigma, \alpha; y) &= -\frac{n}{2}\{(\ln\sigma^2) + \ln(2\pi)\} + \ln|S| - \frac{1}{2\sigma^2}(y'S'MSy) && (8) \\
\ln L(\alpha; y) &= \kappa + \ln|S| - (n/2)\ln(y'S'MSy) && (9)
\end{aligned}
$$

Where κ represents a scalar constant and both $M = I_n - H$ and $H = X(X'X)^{-1}X'$ are idempotent matrices. The term $|S|$ is the Jacobian of the transformation from y to Sy.

Property 3 above allows us to greatly simplify the MESS log-likelihood using:

$$
\begin{aligned}
\mathrm{trace}(D) &= 0 \\
|e^{\alpha D}| &= e^{\mathrm{trace}(\alpha D)} \\
&= e^0 = 1
\end{aligned}
$$

This results in a concentrated log-likelihood taking the form:

$$
\ln L(\alpha; y) = \kappa - (n/2)\ln(y'S'MSy) \qquad (10)
$$

Maximizing the concentrated log-likelihood (10) is equivalent to minimizing $(y'S'MSy)$, the overall sum-of-squared errors with respect to α. Thus, one can interpret the search for an optimal S as a search for a coordinate system (possibly oblique) which has the same multidimensional volume as the orthogonal Cartesian coordinate system that has the same determinant, but yields a better goodness-of-fit or sum-of-squared errors among the variables.

2.4 A closed form solution for the parameters

In contrast to conventional spatial autoregressive models, the MESS model has a closed-form solution. By way of preliminaries, we note that for row-stochastic, non-negative matrices D, a maximum value of 1 will exist in any row, meaning that elements of the powers of D do not rise with the power, so a power series converges rapidly. This allows us to implement the infinite series definition of S using a truncated power series expansion containing q terms. This leads to D^{q-1} as the highest degree term, allowing formation of an nxq matrix Y comprised of powers of D times y as shown in (11).[2]

$$Y = [y \; Dy \; D^2 y \ldots D^{q-1} y] \tag{11}$$

Note that one does not need to compute the n by n matrix S separately, as S always appears in conjunction with the n by 1 vector y.[3] This allows computation of Sy in $O((q-1)n^2)$ operations for dense D by sequential left-multiplication of y by D to form $n-$element vectors, (i.e., Dy, $D(Dy) = D^2 y$, and so on). Typically, the matrix D is sparse when populated with non-zero elements based on a finite number m of nearest neighbors, leading to a dramatic decline in the number of operations required to compute Sy. The number of operations required drops to $O((q-1)n_{\neq 0})$, where $n_{\neq 0}$ denotes the number of non-zeros. For the nearest neighbor spatial weight matrix approach, the operation count associated with computing Sy is linear in n.

To solve for parameter estimates of the model, we define the diagonal q by q matrix W containing part of the coefficients of the power series as shown in (12).

$$W = \begin{pmatrix} 1/0! & & & \\ & 1/1! & & \\ & & \ddots & \\ & & & 1/(q-1)! \end{pmatrix} \tag{12}$$

In addition, we define the q-element column vector v shown in (13) that contains powers of the scalar real parameter α, $|\alpha| < \infty$.

$$v = [1 \; \alpha \; \alpha^2 \ldots \alpha^{q-1}]' \tag{13}$$

Using (11), (12), and (13), we can rewrite Sy as shown in (14).

$$Sy = YWv \tag{14}$$

Pre-multiplying Sy by the least-squares idempotent matrix M yields the residuals e, allowing us to express the overall sum-of-squared errors, $u'u$ as in (15),

$$u'u = v'WY'M'MYWv = v'(WY'MYW)v = v'Qv \tag{15}$$

[2]Given the rapid decline in the coefficients in the power series, $\alpha^t/t!$, achieving a satisfactory progression with nine or ten terms seems feasible. We show later that α takes on values between -2 and 1 for spatial dependence magnitudes encountered in applied practice. For $\alpha = -2$, we have: $(\alpha^9/9!) = -0.0014$ and $(\alpha^{10}/10!) = 0.00028$, which would be multiplied times elements of D that are less than unity.

[3]Sidje (1998) has also used this as a point of departure in the computation of matrix exponentials. In addition, Sidje provides other algorithms for computing the matrix exponential right multiplied by a vector.

where $Q = WY'MYW$. Appendix A contains a proof that their exists an unique minimum to this quadratic form despite the presence of polynomial constraints as embodied in (13).

Turning attention to actually finding the optimum, we write the overall sum-of-squared errors $v'Qv$ as a $2q-2$ degree polynomial in the variable α. The coefficients in the polynomial are the sum of all terms appearing in Q associated with each power of α. The number of coefficients of a $2q-2$ degree polynomial equals $2q-1$ due to the constant term (coefficient associated with the degree 0 term). Specifically, the coefficients c, a $2q-1$ element column vector are shown in (16),

$$c_{t-1} = \sum_{i=1}^{q} \sum_{j=1}^{q} Q_{ij} Ind((i+j) = t) \tag{16}$$

where $Ind()$ is an indicator function taking on values of 1 when the condition is true. The terms associated with the same power of α have subscripts i, j that sum to the same value. For example, $\alpha^i \alpha^j = \alpha^t$ when $i + j = t$, which means that each coefficient c_i is the sum of the elements along the anti-diagonals of Q. This allows us to rewrite $v'Qv$ as the $2q-2$ degree polynomial $P(\alpha)$, shown in (17).

$$P(\alpha) = \sum_{i=1}^{2q-1} c_i \alpha^{i-1} = v'Qv \tag{17}$$

To find the minimum of the sum-of-squared errors, we differentiate the polynomial $P(\alpha)$ in (17) with respect to α, equate to zero, and solve for α as shown in (18).

$$\frac{dP(\alpha)}{d\alpha} = \sum_{i=2}^{2q-1} c_i (i-1) \alpha^{i-2} = 2v'Q \left(\frac{dv}{d\alpha} \right) = 0 \tag{18}$$

The derivative $dP(\alpha)/d\alpha$ is a degree $2q-3$ polynomial and thus has $2q-3$ possible roots. The problem of finding all the roots of a polynomial has a well-defined solution. Specifically, the roots equal the eigenvalues of the companion matrix associated with the polynomial (Horn and Johnson (1993, p. 146-147)).[4] We note that computation of the eigenvalues requires $O(8q^3)$ operations in this case and does not depend upon n, making this approach amenable to large spatial data samples. Solution for the maximum likelihood estimates involves trivial numerical calculations to find eigenvalues of the small companion matrix, so we refer to this as a closed-form solution.

In Appendix A we show that this solution that minimizes the sum-of-squared errors and maximizes the MESS likelihood is unique. Such unique optima do not always occur in spatial statistics. See Warnes and Ripley (1987) and Mardia and Watkins (1989)) for a discussion of the potential multimodality of the likelihood.

2.5 Estimates of dispersion

Given that the log-likelihood can be evaluated rapidly owning to the elimination of the log-determinant term that appears in conventional SAR models, a numerical Hessian approach

[4]Other methods also exist for finding the roots of polynomials. See Press et al. (1996, p. 362-372) for a review of these.

can be used to produce estimates of dispersion for the parameters. We have compared estimates of dispersion from this approach to those from Bayesian estimation in numerous applied settings and found these to be accurate. We note that due to the properties of the matrix exponential, analytical derivatives of the likelihood function are reasonably easy to calculate. This would allow the possibility of a method of moments estimation scheme as in Kelejian and Prucha (1998, 1999), a subject for future investigation.

An alternative approach to inference is to carry out likelihood ratio tests that reflect exclusion of each explanatory variable in the model. Conventional likelihood ratio tests would compute the deviance (twice the difference in the log-likelihoods) and this would asymptotically follow a $\chi^2(J)$ distribution, where J is the number of exclusionary restrictions (Fan, Hung, and Wong (2000)). The signed root deviance equals the square root of twice the difference in log-likelihoods (deviance) between the unrestricted and restricted models, given the sign of the parameter estimate, and can be treated similar to a $t-$statistic for inference (Chen and Jennrich (1996)). A test for significant spatial dependence can be based on the null hypothesis: $\alpha = 0$, which involves using the least-squares log-likelihood in the signed root deviance calculation to produce an inference.

Efficient computation of likelihood ratio tests for individual explanatory variables requires updating the sum-of-squared errors matrix Q without recomputing the actual regressions. Let $\hat{B} = (X'X)^{-1}X'Y$ denote the k by q matrix of estimates from the regression of Y on X, where Y is defined in (11). Let $\hat{E} = Y - X\hat{B}$ denote the n by q matrix of errors from the regression. Expression (19) shows the restricted least squares estimate for $\tilde{B}_j, (j = 1 \ldots q)$,

$$\tilde{B}_j = \hat{B}_j + (X'X)^{-1}R'(R(X'X)^{-1}R')^{-1}(r - R\hat{B}_j) \tag{19}$$

where r and R denote an h by 1 vector, and h by k matrix, constructed to impose h hypotheses.[5]

Let ΔB_j in (20) denote the change in the restricted least squares estimates versus the unrestricted estimates for the jth regression.

$$\Delta B_j = \tilde{B}_j - \hat{B}_j = (X'X)^{-1}R'(R(X'X)^{-1}R')^{-1}(r - R\hat{B}_j) \tag{20}$$

The inner product of any two vectors of restricted regression errors appears in (21).

$$\tilde{E}'_{j1}\tilde{E}_{j2} = (Y - X\tilde{B}_{j1})'(Y - X\tilde{B}_{j2}) = (\hat{E}_{j1} - X\Delta B_{j1})'(\hat{E}_{j2} - X\Delta B_{j2}) \tag{21}$$

where $\tilde{E}_{j1}, \tilde{E}_{j2}$ represent the vectors of restricted regression errors and $j1, j2 = 1, \ldots, q$. Expanding (21) yields (22).

$$\tilde{E}'_{j1}\tilde{E}_{j2} = \hat{E}'_{j1}\hat{E}_{j2} + (\Delta B_{j1})'(X'X)(\Delta B_{j2}) \tag{22}$$

Two of the possible terms vanish due to the enforced orthogonality between the residuals and the data in least-squares, and we can expand the second term from (22) and cancel terms leading to a simple expression shown in (23) for the increase in error arising from restrictions.

[5]See Gentle (1998, p. 166) for the standard restricted least squares estimator as well as some other techniques for computing these estimates.

$$(\Delta B_{j1})'(X'X)(\Delta B_{j2}) = (r - R\hat{B}_{j1})'[R(X'X)^{-1}R']^{-1}(r - R\hat{B}_{j2}) \tag{23}$$

Finally, define the q by q matrix of cross-products of restricted least-squares regressions as $\tilde{E}'\tilde{E}$ with $j1, j2$th element $\tilde{E}'_{j1}\tilde{E}_{j2}$, leading to a restricted sum of squared errors, $Q_R = W(\tilde{E}'\tilde{E})W$.

We note that the quantities $\hat{B}_j$ and the Cholesky factors of $X'X$ are already known from the unrestricted regressions. However, $[R(X'X)^{-1}R']^{-1}$ requires $O(h^3)$ operations for its decomposition. Typically, h will be small, as in the case of testing the effect of deleting a single variable where $h = 1$. Since computing the increase in errors from the restrictions requires $O(h^3)$ operations and resolving the first order conditions requires $O(8q^3)$ operations, deviance (i.e., likelihood ratio) tests or signed root deviance statistics do not depend upon n.

2.6 Spatial model diagnostics

Christensen, Johnson, and Pearson (1992), Haining (1994), and Martin (1992) investigated regression-type diagnostics for spatial models that employ an estimated variance-covariance matrix, $\Omega(\theta)$ parameterized by a vector θ.

Maximum likelihood estimation of SAR models encounters difficulties with many of the standard deletion diagnostics, which often rely upon linearity for speed. The logistical difficulties of performing n non-linear optimizations to perform deletion diagnostics poses a barrier to the routine usage of these popular diagnostics in a spatial setting. However, the simplicity and speed of MESS can facilitate computation of these diagnostics in a spatial setting. As an example, if we let $Z = MYW$, so that $Q = Z'Z$, and $z_i = (Z'_{ij}, j = 1, \ldots, q)$, a q by 1 vector. We denote Q when the ith observation is deleted as $Q_{(i)}$ and rely on standard regression results for one-out sum-of-squared errors (e.g., Christensen (1996, p. 345)), as in (24).

$$Q_{(i)} = Q - \frac{z_i z_i'}{(1 - H_{ii})} \tag{24}$$

Using the same quick mechanism for finding roots of polynomials we can produce a sequence of row-deletion diagnostic statistics that measure the influence or leverage impact of excluding individual observations. Of course, these diagnostics might be altered to delete a single observation as well as neighboring observations in a spatial setting as pointed out by Christensen (1996, p. 348).

3 Bayesian estimation of the model

A Bayesian approach to the MESS model includes specification of prior distributions for the parameters α, β, σ in the model. Prior information regarding the parameters β and σ is unlikely to exert much influence on the posterior distribution of these estimates in the case of very large samples, where the MESS model holds an advantage over more traditional spatial autoregressive models. However, priors placed on the parameter α may exert an influence even in large samples, because of the important role played by spatial dependence in these models.

In section 3.1 we develop Bayesian estimates for the canonical case of a normal-gamma prior, that is a normal for β and inverted gamma prior for σ, as well as an arbitrary prior for α. Univariate numerical integration is required with respect to the parameter α, so we have flexibility in the choice of this prior and proceed using $\pi(\alpha)$ to denote an arbitrary prior. When these priors for β, σ, α become diffuse, the log-marginal posterior for α approaches an expression proportional to the log-likelihood concentrated with respect to β and σ, indicating that the posterior mode from Bayesian estimation which involves univariate numerical integration over α would equal the maximum likelihood estimate for α. This is a result equivalent to that found for Bayesian estimation of traditional regression models.

3.1 The case of normal-gamma conjugate priors

The prior pdf's for the parameters β and σ take the normal, inverted gamma conjugate forms shown in (25), and we let $\pi(\alpha)$ denote an arbitrary prior.

$$\begin{aligned}
\pi(\beta,\sigma) &= \pi(\beta|\sigma)\pi(\sigma) \\
\pi(\beta|\sigma) &= \frac{|C|^{1/2}}{(2\pi)^{k/2}\sigma^k}\exp\left[-\frac{1}{2\sigma^2}(\beta-\bar{\beta})'C(\beta-\bar{\beta})\right] \\
\pi(\sigma) &= \frac{K}{\sigma^{\nu+1}}\exp\left(-\frac{\nu\bar{s}^2}{2\sigma}\right) \\
K &= 2(\nu\bar{s}^2/2)^{\nu/2}/\Gamma(\nu/2) \qquad (25)
\end{aligned}$$

The normal prior pdf for β has prior mean $\bar{\beta}$ and variance-covariance matrix $\sigma^2 C$, where C is a positive definite symmetric matrix specified by the practitioner. If we let $C = [(1/g)X'X]^{-1}$, a kxk matrix based on the sample data matrix X and the scalar g which controls prior uncertainty, we have the $g-$prior proposed by Zellner (1986). The inverted gamma prior pdf for σ has parameters ν and $\bar{s}^2$, where $0 < \nu, \bar{s}^2 < \infty$ for a proper prior. K is a normalizing constant, and Γ is the gamma function.

Using Bayes' theorem to combine the likelihood and prior, we will be interested in the marginal posterior for the parameter α, which can be obtained by analytically integrating out the elements of β and the parameter σ (Zellner, 1971 pp. 308-09). Specifically:

$$\begin{aligned}
& \int \pi(\beta|\sigma)\pi(\sigma)\pi(\alpha)p(y|\beta,\sigma,\alpha)d\beta d\sigma d\alpha \qquad (26) \\
= \; & \frac{K|C|^{1/2}}{2\pi^{(n+k)/2}}\int \frac{1}{\sigma^{n+\nu+k+1}}\exp(-\frac{1}{2\sigma^2}[\nu\bar{s}^2 + P(\alpha) \\
+ \; & (\beta-\bar{\beta})'C(\beta-\bar{\beta}) + (\beta-\hat{\beta})'X'X(\beta-\hat{\beta})])\pi(\alpha)d\beta d\sigma d\alpha
\end{aligned}$$

where $\hat{\beta} = (X'X)^{-1}X'S(\alpha)y$, regression estimates for β assuming α is known, and $P(\alpha)$ is the $2q-2$ degree polynomial shown in (17). Recall that the determinant $|S|$ will be unity by virtue of Property 3, so this term does not appear in (26). Using the properties of the multivariate normal pdf and the inverted gamma pdf to integrate with respect to β and σ, we define:

$$\begin{aligned} Q_1 &= (\bar{\beta} - \tilde{\beta})' C (\bar{\beta} - \tilde{\beta}), \\ Q_2 &= (\hat{\beta} - \tilde{\beta})' X'X (\hat{\beta} - \tilde{\beta}), \\ \tilde{\beta} &= A^{-1}(C\bar{\beta} + X'X\hat{\beta}), \\ A &= X'X + C, \end{aligned}$$

and arrive at:

$$\begin{aligned} p(\alpha|y) &= \frac{1}{2}\frac{K}{(2\pi)^{n/2}}\left(\frac{|C|}{|A|}\right)^{1/2} 2^{(n+\nu)/2}\Gamma[(n+\nu)/2] \\ &\cdot \int \left(\nu\bar{s}^2 + P(\alpha) + Q_1 + Q_2\right)^{-(n+\nu)/2} \pi(\alpha)d\alpha \end{aligned} \tag{27}$$

Note that the determinants $|A|$ and $|C|$ do not depend on α, and $P(\alpha) = \sum_{i=1}^{2q-1} c_i\alpha^{i-1}$. We can define a constant of integration:

$$\kappa = \frac{1}{2}\frac{K}{(2\pi)^{n/2}}\left(\frac{|C|}{|A|}\right)^{1/2} 2^{(n+\nu)/2}\Gamma[(n+\nu)/2] \tag{28}$$

allowing further simplification of (27).

$$p(\alpha|y) = \kappa \int \left(\nu\bar{s}^2 + \sum_{i=1}^{2q-1} c_i\alpha^{i-1} + Q_1 + Q_2\right)^{-(n+\nu)/2} \pi(\alpha)d\alpha \tag{29}$$

We consider the case where the priors for β, σ and α become diffuse. For the prior on β, we can let the roots of $C \rightarrow 0$ leading to a null matrix C, or let $g \rightarrow \infty$ in the case of the $g-$prior. These lead to $Q_1 \rightarrow 0$ and $\hat{\beta} \rightarrow \tilde{\beta}$, meaning the term $Q_2 \rightarrow 0$. The prior on σ becomes diffuse when $\nu, \bar{s}^2 \rightarrow 0$, eliminating the term $\nu\bar{s}^2$ and creating an exponent of $-n/2$. Assuming the arbitrary prior on α is diffuse, the expression in (29) approaches:

$$p(\alpha|y) \propto \int \left(\sum_{i=1}^{2q-1} c_i\alpha^{i-1}\right)^{-n/2} d\alpha \tag{30}$$

Univariate numerical integration can be carried out to find the posterior distribution for α in the case of (29) with priors, or in the case of (30) where the priors are diffuse. We note that in the face of diffuse priors, the log of $p(\alpha|y)$ in (30) is proportional to the log-likelihood function for the MESS model, since $P(\alpha) = \sum_{i=1}^{2q-1} c_i\alpha^{i-1} = (y'S'MSy)$ as in (10). The means that numerical integration over α will produce a posterior mode for α that will equal the value from maximizing the likelihood over α, a result similar to that found for Bayesian regression analysis.

It is also the case that as n the number of observations grows large, the expression in (29) approaches (30), a point made by Jeffreys and developed fully in Zellner (1971, pp. 31-34). This follows from the fact that the log-likelihood will be of order n, whereas prior information does not depend on the sample size n, leading the likelihood factor to dominate the posterior pdf in (29).

It is informative to contrast this result with that arising in SAR models. Following a similar approach to that used above where the priors on β, σ and ρ become diffuse, we arrive at the marginal posterior for ρ, shown in (31) (see Hepple (1995a,1995b)).

$$\begin{aligned} p(\rho|y) &\propto \int |I_n - \rho D| \, (u'u)^{-n/2} \, d\rho \\ u &= (I_n - \rho D)y - X\beta(\rho) \\ \beta(\rho) &= (X'X)^{-1}X'(I_n - \rho D)y \end{aligned} \tag{31}$$

The univariate numerical integration over ρ ranges in the open interval $(1/\mu_{min}, 1/\mu_{max})$, where μ are eigenvalues of the weight matrix D.[6] This requires calculating the eigenvalues of the n by n spatial weight matrix as well as the log-determinant of the n by n matrix $(I_n - \rho D)$ over this range of ρ values, a computationally expensive operation.[7]

In contrast, solution for α in the case of diffuse priors for the Bayesian MESS model requires simple univariate integration involving the scalar polynomial $P(\alpha)$. For the case involving informative priors in (29), we also need to include the prior $\pi(\alpha)$ as well as $\hat{\beta}(\alpha) = (X'X)^{-1}X'S(\alpha)y$ and $\tilde{\beta}(\alpha) = A^{-1}(C\bar{\beta} + X'X\hat{\beta}(\alpha))$ in the univariate numerical integration procedure.

Turning attention to the posterior distribution for β in the Bayesian MESS model, we can use the multivariate $t-$density centered at $\beta(\alpha)$, suggesting that the posterior mean can be computed analytically using:

$$E(\beta|y, X) = (X'X + C)^{-1}(X'S(\bar{\alpha})y + C\bar{\beta}) \tag{32}$$

where $\bar{\alpha}$ denotes the posterior mean from (29). As in the case of α, diffuse priors on β lead to C approaching a null matrix, producing a posterior mode for the β parameters that will equal those from maximum likelihood. The posterior covariance matrix unconditional on α takes the form shown in (33), which requires univariate integration of the scalar polynomial expression for $P(\alpha)$.

$$\text{var-cov}(\beta) = \frac{1}{n+\nu}\left(\int (P(\alpha)p(\alpha|y, X)\pi(\alpha)d\alpha\right)(X'X + C)^{-1} \tag{33}$$

Given the multivariate $t-$density for β, individual elements β_i take the form of a univariate Student $t-$density conditional on α, and are essentially the standard expressions for normal multiple regression from Zellner (1971, pp. 67-68). Here as in the case of the posterior distribution for α, the scalar polynomial expression $P(\alpha)$ plays an important role in simplifying the tasks involved.

As already noted, we use univariate integration to compute: $E(P(\alpha)|y, X) = \bar{\sigma} = \int P(\alpha)p(\alpha|y, X)\pi(\alpha)d\alpha$ as part of our solution for the variance-covariance matrix for β. This provides us with an expression for the posterior mean of σ.

[6] See for example Lemma 2 in Sun et al., 1999.

[7] Barry and Pace (1999) provide a computationally efficient approach to estimating the log-determinant over a grid of values for $\rho \in [1/\mu_{min}, 1/\mu_{max}]$. In addition, they provide a vector expression for $-n/2 \cdot log(u'u)$ ranging over values of ρ in this grid which also facilitates univariate numerical integration.

4 Bayesian model comparison

Other authors have set forth the Bayesian theory behind model comparison which involves specifying prior probabilities for each of the discrete set of alternative models $M = M_1, M_2, \ldots, M_m$ under consideration, which we label $\pi(M_i)$, as well as prior distributions for the parameters $\pi(\eta)$, where $\eta = (\alpha, \beta, \sigma)$ (e.g., Zellner, 1971 and Fernandez, Ley, and Steel, 2001b). Our focus here is on two types of model comparison. In section 4.1 we consider models using alternative spatial weight matrices, and in section 4.2 we consider models with different explanatory variables. We suggest alternative priors for $\pi(\eta)$ for these two scenarios in sections 4.1 and 4.2.

If the sample data are to determine the model probabilities, the prior probabilities should be set to equal values of $1/m$, making each model equally likely apriori. These are combined with the likelihood for y conditional on η as well as the set of models M, which we denote $p(y|\eta, M)$. The joint probability for M, η, and y takes the form:

$$p(M, \eta, y) = \pi(M)\pi(\eta|M)p(y|\eta, M) \tag{34}$$

Application of Bayes rule produces the joint posterior for both models and parameters as:

$$p(M, \eta|y) = \frac{\pi(M)\pi(\eta|M)p(y|\eta, M)}{p(y)} \tag{35}$$

The posterior probabilities regarding the models takes the form:

$$p(M|y) = \int p(M, \eta|y) d\eta \tag{36}$$

which requires integration over the parameter vector η. Although the Bayesian theory of model choice is elegant, integration over the parameter space represents one problem that arises when implementing this approach. As already shown, obtaining the marginal posterior for the MESS model is relatively simple in comparison with traditional spatial autoregressive models.

A second problem that arises in Bayesian model comparison is that posterior model probabilities can be sensitive to specification of the prior information. Use of diffuse priors on the model parameters might seem desirable in this situation, but can lead to paradoxical outcomes as noted by Lindley (1957). We suggest one approach to dealing with this problem for the case of model comparisons involving alternative spatial weight matrices. In section 4.2 where we deal with model comparison involving alternative explanatory variables, we draw on the work of Fernandez, Ley, and Steel (2001b) to resolve this type of problem. This second type of model comparison employs a Markov chain Monte Carlo composition method often labelled MC^3. The MC^3 method requires posterior odds ratios that relate model i to model j as shown in (37).[8]

$$O_{i,j} = \frac{p(M = i|y)}{p(M = j|y)} \tag{37}$$

[8]In the case of uniform prior model probabilities, these odds ratios are referred to as Bayes factors.

$$= \frac{\pi(M=i)\int \pi(\eta|M_i)p(y|\eta, M_i)d\eta}{\pi(M=j)\int \pi(\eta|M_j)p(y|\eta, M_j)d\eta}$$

In the context of regression models, Zellner (1971, pp. 310-11) considers the odds ratio for the special case of large samples with identical inverted gamma priors for the parameters σ, and identical normal priors for β in both models. He shows that in this special case, the odds ratio approaches a likelihood ratio test, depending exclusively on the relative fit of the two models. This type of situation could arise when comparing models i and j based on different spatial weight matrices, where all other aspects of the model specification are held constant. A similar result occurs for the MESS model by adding to Zellner's development an identical normal prior on α for both models.[9] This approach yields a Bayes factor constructed using the univariate integral expression (30) in both the numerator and denominator. As already noted, the integral is analogous to the polynomial expression in α that appears in the concentrated log-likelihood function representing overall model fit.

4.1 Model comparison over alternative spatial weight matrices

Economic theory often suggests the existence of externalities, spillovers, and other spatial phenomenon, but rarely suggests the exact extent of the dependence. Typically, estimates based on a small number of candidate weight matrices are compared with a final set of estimates and inferences based on one of these spatial weight matrices. Bayesian model comparison allows a formal treatment of this type of specification uncertainty. Estimates and inferences can be based on results averaged across models consisting of alternative weight matrices, referred to as Bayesian model averaging. Posterior model probabilities are used to produce a single set of estimates that represent a linear combination constructed using model probabilities as weights. For simplicity, consider a set of possible models formed using spatial weight matrices based on 1 to m nearest neighbors. Evaluating the log-marginal for a vector of values of α over a range of support ($[\alpha_{\min}, \alpha_{\max}]$) yields a vector of log-marginal values that collectively provides a finite representation of the log-marginal density. Given this vector, numerical integration can proceed using a variety of different methods such as Simpson's rule.[10]

The key quantity for model comparison is often referred to as the "integrated likelihood," which we designate as L^I. The integrated likelihood for MESS model comparison would likely represent a large sample situation equivalent to diffuse priors on β and σ. This produces an expression involving only α, which will be proper if and only if the product of the integrated likelihood and prior $\pi(\alpha)$ are bounded, that is:

$$0 < \int_{-\infty}^{+\infty} L^I(\alpha; y)\pi(\alpha)d\alpha < \infty \tag{38}$$

As already noted in discussion of maximum likelihood estimation, a unique interior α maximizes the likelihood (concentrated with respect to β and σ). The issue of posterior

[9] We motivate choice of a normal prior more fully in the next section.

[10] Integration involves the anti-log which often requires scaling. Effective scaling can be accomplished by storing the vectors of log-marginals as columns in a matrix and subtracting the maximum matrix element from all elements in the matrix. This produces a value of zero as the largest element, with an anti-log of unity, providing a solution to the scaling problem that requires no tuning.

propriety depends upon specific priors for the spatial dependence parameter α. We advocate use of a normal prior with mean zero and a standard deviation assigned by the investigator to reflect uncertainty regarding this parameter. This choice of prior is motivated by a number of considerations. First, since $-\infty < \alpha < \infty$ and the matrix exponential spatial structure exhibits symmetry, a normal prior on the strength of spatial dependence seems reasonable. This is in contrast to SAR models where in general the spatial dependence parameter ρ has support over a bounded, asymmetric region.

A second motivation for a normal prior on α is that in large samples the posterior for α should approach a normal distribution. This stems from the argument in Zellner (1971, pp. 31-34), where the posterior distribution for any Bayesian model can be expanded around the maximum likelihood estimate using a Taylor series. The first term of this expansion represents a normal distribution that can be used with accuracy of $1/\sqrt{n}$ to approximate the posterior. Therefore, a normal prior for α serves as a pseudo conjugate prior in applied problems with large n. To illustrate this point Figure 2 presents the posterior distribution for α from a data-generated example based on a sample of $n = 3,107$ US county observations and diffuse priors for all parameters in the model. This posterior is displayed alongside a normal distribution with the same mean and variance, and these show a close correspondence between the posterior and normal distributions.

Specification of prior uncertainty regarding α is facilitated by the correspondence between the spatial dependence parameter α and the spatial autocorrelation parameter $\rho \doteq 1 - exp(\alpha)$ from the SAR model. This can be used to construct an interval $(\alpha_{min}, \alpha_{max})$ where α has most of its prior support, allowing a prior standard deviation to be formulated.

Alternative approaches to developing priors for α in situations involving model comparison would be to develop Jeffreys' independence prior, or a reference prior of the type set forth in Berger and Bernardo (1989). These types of priors have advantages over informative priors (see Berger, Pericchi (1996a,1996b) and Berger, Pericchi, and Varshavsky, 1998). The literature also contains suggestions for a 'deviance information criterion' (Speigelhalter et al. 2001) that combines measures of model fit with a penalty for model complexity based on the number of parameters. We do not pursue these approaches here, but a subject for future work would be to examine implementing these methods in the context of Bayesian MESS model comparison.

Identical priors for the parameters β, σ seem reasonable for the case where the models compared differ only with respect to the spatial weight matrices employed. As already noted in this special case, a large sample results in Bayes factors that approach a likelihood ratio test. We note that the parameters $\eta = (\beta, \sigma)$ of the MESS model would not require explicit estimation to conduct model comparison tests in this situation. The ability to analytically integrate out the parameters β and σ serves the same role as concentrating the log-likelihood with respect to these parameters in a maximum likelihood setting. Computation of Bayes factors would only require univariate integration over α. Of course, a proper prior on α must be used and sensitivity of the inferences to this prior need to be checked. There is an intuitive appeal to the notion that in large samples posterior model inferences should be robust with respect to prior information. Despite this intuitive appeal, this should be explored in particular applications. The computational speed of MESS allows practitioners to construct a range of posteriors based on varying priors and explore the robustness of posterior model inferences. We will illustrate posterior model comparison of alternative spatial weight matrices in section 5, and illustrate the role of prior information.

4.2 Model comparisons involving alternative explanatory variables

Raftery, Madigan, and Hoeting (1997), Fernandez, Ley, and Steel (2001b), and others have written extensively on Bayesian model averaging over alternative linear regression models containing differing explanatory variables. The Markov Chain Monte Carlo model composition (MC^3) approach introduced in Madigan and York (1995) is set forth here for the MESS model. For a regression model with an intercept and k possible explanatory variables, there are 2^k possible ways to select regressors to be included or excluded from the model. Later, we consider an example with $k = 21$ which results in $2,097,152$ possible models, and computation of the log-marginal for all possible models quickly becomes tedious.

The MC^3 method of Madigan and York (1995) devises a strategic stochastic process that can move through the potentially large model space and sample regions of high posterior support. This eliminates the need to consider all models by constructing a sample from relevant parts of the model space, while ignoring irrelevant models. Specifically, they construct a Markov chain $M(i), i = 1, 2, \ldots$ with state space Ξ that has an equilibrium distribution $p(M_i|\theta)$, where $p(M_i|\theta)$ denotes the posterior probability of model M_i based on the data θ. This Markov chain is simulated for $i = 1, \ldots, T$, which will converge as $T \to \infty$ almost surely to $E(p(\eta|M, \theta)$ under certain regularity conditions (Smith and Roberts, 1993). The Markov chain is based on a neighborhood, nbd(M) for each $M \in \Xi$, which consists of the model M itself along with models containing either one variable more, or one variable less than M. The addition of an explanatory variable to the model is often labelled a 'birth process' whereas deleting a variable from the set of explanatory variables is called a 'death process'. A transition matrix, s, is defined by setting $s(M \to M') = 0$ for all $M' \ni \text{nbd}(M)$ and $s(M \to M')$ constant for all $M' \in \text{nbd}(M)$. If the chain is currently in state M, we proceed by drawing M' from $s(M \to M')$. This new model is then accepted with probability:

$$\min\left[1, \frac{p(M'|y)}{p(M|y)}\right] = [1, O_{M',M}] \tag{39}$$

Where $O_{M',M}$ is the odds ratio set forth in (37). Note, the computational ease of constructing odds ratios (or Bayes factors for the case of equal prior probabilities assigned to all candidate models) facilitates construction of a Metropolis-Hastings sampling scheme for the MC^3 method. A vector of the log-marginal values for the current model M is stored during sampling along with a vector for the proposed model M'. These are then scaled and integrated to produce $O_{M',M}$ which is used in (39) to determine whether to accept the new model or stay with the current model. Saving the log-marginal density vectors for each unique model found during the MCMC sampling allows calculation of posterior model probabilities over the set of all unique models visited by the sampler.[11]

[11] Although the use of birth and death processes in the context of Metropolis-Hastings sampling will theoretically produce samples from the correct posterior, Richardson and Green (1997), among others, advocate incorporating a 'move step' in addition to the birth and death steps into the algorithm. There is evidence that combining these move steps that keep the dimension fixed aid convergence of the sampling process (Denison et al. 1998, Richardson and Green, 1997). The move step takes the form of replacing a randomly chosen single variable in the current explanatory variables matrix with a randomly chosen variable

Specifically, we follow Fernandez, Ley and Steel (2001a) and employ Zellner's $g-$prior (Zellner, 1986) for the parameters β in the model:

$$Sy = \iota\beta_0 + X\beta + \varepsilon \tag{40}$$

where the k the columns of X are in deviations from their means, so that $\iota' X = 0$. The $g-$prior on the regression coefficients β_{M_i} takes the form shown in (41), and we can rely on setting $g = 1/\max\{n, k^2\}$ as motivated by Fernandez, Ley and Steel (2001a, 2001b).

$$\begin{aligned} \beta_{M_i}|\sigma, \rho &\sim N[0, \sigma^2((1/g)X'_{M_i}X_{M_i})^{-1}] \\ \pi(\alpha, \sigma) &\propto 1/\sigma \end{aligned} \tag{41}$$

The marginal likelihood $p(y|M = i)$ under model M_i, the $g-$prior, and uniform priors across models takes the form:

$$\int (\frac{1}{\sqrt{n}} \frac{\Gamma((n-1)/2)}{2\pi^{(n-1)/2}(1+g)^{k_i/2}} \left[y'S'(I - H_\iota - \frac{g}{1+g}H_{M_i})Sy \right]^{-(n-1)/2} d\alpha \tag{42}$$

where H_ι is the projection matrix arising from the constant parameter β_0, and H_{M_i} is the projection matrix associated with model M_i containing k_i explanatory variables. Again, the MESS univariate polynomial expression can be used in this setting allowing integration of the log-marginal likelihood during MCMC sampling.

5 Applied illustrations

The first application in section 5.1 illustrates the ability of the MESS model to produce estimates and inferences similar to those from conventional SAR models. We have pointed to numerous analytical advantages of the MESS model over conventional SAR models. Here we illustrate that this model can provide an alternative to SAR models in applied problems. A second illustration in section 5.2 demonstrates that prior information has a relatively small impact for problems involving moderately large sample sizes. The third application in section 5.3 illustrates model comparison where the models differ in terms of the spatial weight matrix employed. A final application in section 5.4 considers model comparison in the context of MC^3 determination of appropriate explanatory variables, where models differ in terms of the explanatory variables.

5.1 A comparison of conventional and MESS models

Information from the 1997 Census of Agriculture on a sample of 24,473 agricultural zip code areas was used to consider the impact of the conservation reserve program (CRP) and

---

not currently in the model. Specifically, proposing a model with one less explanatory variable (death step) and then adding an explanatory variable to this new model proposal (birth step) leaves the resulting model proposal with the same dimension as the original one, but with a single component altered. This type of sampling process is often labelled 'reversible jump' MCMC. The model proposals that result from birth, death and move steps are all subjected to the Metropolis-Hastings accept/reject decision shown in (39), which is valid so long as the probabilities of birth, death and move steps have equal probability of 1/3.

wetlands reserve program (WRP) on acres harvested in each zip code area. The dependent variable is the log of acres harvested. Explanatory variables included various categories of agricultural land-use measured by logged acres in: land in CRP and WRP conservation programs, idle land, pasture land, rangeland, woodland, land under soil improvement programs, acres of failed crops, fallow land and an other-land category. Additional explanatory variables were: the number of farms, the number of various livestock (e.g., beef and milk cows, hogs and pigs, sheep and lambs, hens and pullets, horses and ponies), a variable indicating the proportion of owner-operated farms, the proportion of population in the zip code area that is classified as rural and the proportion classified as farm population. All of these variables were log-transformed as well.

A spatial lag of each land-usage variable (land in CRP and WRP conservation programs, idle land, pasture land, rangeland, woodland, land under soil improvement programs, acres of failed crops, fallow land and an other-land category) was also included in the model to capture the effect of neighboring zip-code area land-usage on acres harvested. The spatial lag was constructed by multiplying the spatial weight matrix W times the land use variables (e.g., $y = \rho W y + X\beta + W\tilde{X}\theta + \varepsilon$, where $\tilde{X}$ contains the land usage explanatory variables). The spatial weight matrix was constructed using first-order contiguity.

An economic issue of interest is the elasticity response of acres harvested with respect to acreage placed in the CRP and WRP conservation programs. Table 1 presents estimates from the conventional SAR model alongside those from maximum likelihood MESS. In addition to the point estimates, asymptotic $t-$statistics are presented for both models based on a numerical Hessian calculation. The use of $t-$ratios in the table facilitates comparison between the two sets of estimated standard deviations.

Estimates for all 29 explanatory variables exhibit similar magnitudes and standard deviations such that inferences regarding the significance of these variables would be identical from both sets of estimates. This is also true of the spatial dependence parameters α and ρ as well as the noise variance, σ^2, and the R^2 statistics measuring fit of the two models. Note, the correspondence $\rho = 1 - \exp(\alpha)$ suggests a value of $\hat{\rho} = 0.3589$, close to the estimate of 0.41 in the table. Given the estimates in Table 1, both the CRP and WRP conservation programs had a negative and significant effect on acres harvested. Since the coefficient on the spatial lag of conservation acreage is not significant, land placed in the conservation programs in surrounding zip code areas did not exert a significant impact on acreage harvested.

5.2 The impact of prior information and sample size

The impact of using subjective prior information to specify prior means and variances for the parameters β, α, and σ should not exert undue influence on the Bayesian MESS model estimates and inferences in reasonably large spatial samples. Here we provide some evidence regarding the impact of prior information for sample sizes of 3,107 observations and 24,473 agricultural zip code areas from the previous illustration.

Sample data concerning voter participation in the 1980 Presidential election for 3,107 US counties from Pace and Barry (1997) was used in conjunction with three priors as well as maximum likelihood estimation. The model uses the log of voter participation rates as the dependent variable with a constant, median household income, proportion of

Table 1: A comparison of MESS and SAR estimates

Variable	MESS $\hat{\beta}$	MESS t	SAR $\hat{\beta}$	SAR t
constant	-0.1719	-13.14	-0.1844	-14.37
conservation (CRP+WRP)	-0.0343	-12.08	-0.0328	-11.79
idle land	0.0441	13.75	0.0418	13.30
pasture land	0.0771	17.12	0.0760	17.20
rangeland	-0.0490	-12.54	-0.0488	-12.73
woodland	0.0243	6.83	0.0215	6.17
soil improvement	0.0150	4.84	0.0135	4.42
failed acreage	0.0187	6.46	0.0162	5.71
fallow land	0.0364	12.58	0.0349	12.28
other land	0.1663	24.65	0.1699	25.63
farms	0.8613	106.88	0.8438	105.44
beef cows	-0.0466	-12.90	-0.0464	-13.09
milk cows	0.0228	11.21	0.0223	11.21
hogs and pigs	0.0091	3.77	0.0092	3.87
sheep and lambs	0.0129	5.22	0.0129	5.36
hens and pullets	-0.0122	-4.34	-0.0107	-3.88
horses and ponies	-0.0883	-29.26	-0.0843	-28.32
owner operator	-0.0234	-20.30	-0.0218	-19.24
rural population	0.0190	15.13	0.0192	15.65
farm population	-0.0080	-8.49	-0.0075	-8.10
W·(conservation)	-0.0018	-0.45	-0.0044	-1.12
W·(idle land)	-0.0644	-11.55	-0.0542	-9.93
W·(pasture land)	-0.0373	-6.12	-0.0345	-5.76
W·(rangeland)	-0.0644	-12.55	-0.0574	-11.38
W·(woodland)	-0.0057	-1.31	-0.0030	-0.69
W·(soil improvement)	0.0416	7.89	0.0449	8.68
W·(failed acreage)	-0.0084	-1.77	-0.0051	-1.10
W·(fallow land)	0.0220	5.42	0.0191	4.79
W·(other land)	-0.2697	-34.38	-0.3249	-36.27
α	-0.4446	-63.62		
ρ			0.4129	65.42
σ^2	0.0538		0.0518	
R^2	0.9581		0.9457	

Table 2: The impact of prior information for $n = 3,107$ observations

Posterior mean estimates				
variables	tight prior	medium prior	loose prior	maximum likelihood
constant	0.6099	0.6866	0.6963	0.6964
college graduates	0.2520	0.2703	0.2726	0.2726
homeowners	0.4955	0.5048	0.5059	0.5059
income	-0.1000	-0.1253	-0.1286	-0.1286
α	-0.6768	-0.6755	-0.6755	-0.6752
σ^2	0.0154	0.0153	0.0153	0.0153
Posterior standard deviations				
variables	tight prior	medium prior	loose prior	maximum likelihood
constant	0.0397	0.0421	0.0424	0.0435
college graduates	0.0134	0.0139	0.0139	0.0156
homeowners	0.0151	0.0152	0.0152	0.0152
income	0.0156	0.0164	0.0165	0.0171
α	0.0235	0.0235	0.0235	0.0232
σ^2	0.0004	0.0004	0.0004	0.0004

population over 25 years having college degrees, and proportion of population owning homes as explanatory variables. All variables are log-transformed so coefficient estimates reflect elasticity responses. This suggests a prior mean for β of zero may be appropriate. Three alternative diagonal prior covariances were used based on prior standard deviations of 1, 10, and $10,000$ which we refer to as tight, medium and loose imposition of the prior means. For the parameter α a prior mean of zero was used with three alternative prior standard deviations of 1, 3, and 100. The prior parameters $\nu, \bar{s}^2$ for the noise variance were set to 0.01, reflecting a relatively uninformative prior. We might expect the tight prior to produce posterior estimates exhibiting some shrinkage towards the prior mean values of zero for both α and β. The loose prior should exhibit posterior estimates similar to those from maximum likelihood estimation.

Estimation results are presented in Table 2 and the posterior mean estimates associated with all three priors are similar to the point estimates from maximum likelihood estimation. There is some evidence that posterior means associated with tight imposition of the prior produces shrinkage towards zero in the posterior estimates. The table also presents posterior estimates for the standard deviation of the parameters, where again the loose imposition of the prior produces results similar to maximum likelihood.

As the sample size grows, the impact of subjective prior information on the posterior estimates should become smaller. Since the agriculture model from the previous section involved log-transformations with coefficients reflecting elasticity responses, the same three priors were applied to this model to explore this issue.

Posterior mean estimates for this model are presented in Table 3, and posterior standard deviations in Table 4. No difference exists in the first three decimal digits of the posterior estimates associated with the medium and loose prior and those from maximum likelihood. In the case of the tight prior, estimates are the same to two decimal digits. Posterior standard deviations are identical to three or four decimal digits for all three priors and maximum likelihood estimates.

Table 3: Posterior mean estimates for $n = 24,473$ observations

variables	tight prior	medium prior	loose prior	maximum likelihood
constant	-0.1709	-0.1718	-0.1719	-0.1719
conservation (CRP+WRP)	-0.0343	-0.0343	-0.0343	-0.0343
idle land	0.0442	0.0442	0.0441	0.0441
pasture land	0.0773	0.0772	0.0772	0.0772
rangeland	-0.0489	-0.0490	-0.0490	-0.0490
woodland	0.0244	0.0244	0.0244	0.0244
soil improvement	0.0151	0.0151	0.0151	0.0151
failed acreage	0.0188	0.0188	0.0188	0.0188
fallow land	0.0365	0.0365	0.0365	0.0365
other land	0.1667	0.1664	0.1663	0.1663
farms	0.8604	0.8613	0.8614	0.8614
beef cows	-0.0466	-0.0467	-0.0467	-0.0467
milk cows	0.0228	0.0228	0.0228	0.0228
hogs and pigs	0.0092	0.0092	0.0092	0.0092
sheep and lambs	0.0129	0.0129	0.0129	0.0129
hens and pullets	-0.0122	-0.0123	-0.0123	-0.0123
horses and ponies	-0.0883	-0.0884	-0.0884	-0.0884
owner operator	-0.0234	-0.0234	-0.0234	-0.0234
rural population	0.0190	0.0190	0.0190	0.0190
farm population	-0.0080	-0.0080	-0.0080	-0.0080
W·(conservation)	-0.0018	-0.0018	-0.0018	-0.0018
W·(idle land)	-0.0645	-0.0645	-0.0645	-0.0645
W·(pasture land)	-0.0375	-0.0374	-0.0374	-0.0374
W·(rangeland)	-0.0645	-0.0644	-0.0644	-0.0644
W·(woodland)	-0.0059	-0.0058	-0.0058	-0.0058
W·(soil improvement)	0.0416	0.0417	0.0417	0.0417
W·(failed acreage)	-0.0085	-0.0085	-0.0085	-0.0085
W·(fallow land)	0.0220	0.0220	0.0220	0.0220
W·(other land)	-0.2693	-0.2697	-0.2697	-0.2697
α	-0.4446	-0.4446	-0.4446	-0.4446
σ^2	0.0538	0.0538	0.0538	0.0538

Table 4: Posterior standard deviations for $n = 24,473$ observations

variables	tight prior	medium prior	loose prior	maximum likelihood
constant	0.0129	0.0129	0.0129	0.0131
conservation (CRP+WRP)	0.0028	0.0028	0.0028	0.0028
idle land	0.0032	0.0032	0.0032	0.0032
pasture land	0.0045	0.0045	0.0045	0.0045
rangeland	0.0039	0.0039	0.0039	0.0039
woodland	0.0036	0.0036	0.0036	0.0036
soil improvement	0.0031	0.0031	0.0031	0.0031
failed acreage	0.0029	0.0029	0.0029	0.0029
fallow land	0.0029	0.0029	0.0029	0.0029
other land	0.0065	0.0065	0.0065	0.0067
farms	0.0074	0.0074	0.0074	0.0081
beef cows	0.0036	0.0036	0.0036	0.0036
milk cows	0.0020	0.0020	0.0020	0.0020
hogs and pigs	0.0024	0.0024	0.0024	0.0024
sheep and lambs	0.0025	0.0025	0.0025	0.0025
hens and pullets	0.0028	0.0028	0.0028	0.0028
horses and ponies	0.0030	0.0030	0.0030	0.0030
owner operator	0.0011	0.0011	0.0011	0.0012
rural population	0.0013	0.0013	0.0013	0.0013
farm population	0.0009	0.0009	0.0009	0.0009
W·(conservation)	0.0040	0.0040	0.0040	0.0040
W·(idle land)	0.0056	0.0056	0.0056	0.0056
W·(pasture land)	0.0061	0.0061	0.0061	0.0061
W·(rangeland)	0.0051	0.0051	0.0051	0.0051
W·(woodland)	0.0044	0.0044	0.0044	0.0044
W·(soil improvement)	0.0053	0.0053	0.0053	0.0053
W·(failed acreage)	0.0048	0.0048	0.0048	0.0048
W·(fallow land)	0.0041	0.0041	0.0041	0.0041
W·(other land)	0.0065	0.0065	0.0065	0.0078
α	0.0077	0.0077	0.0077	0.0070
σ^2	0.0005	0.0005	0.0005	0.0005

Table 5: Posterior model probabilities for alternative weights and priors

Neighbors/priors	$g = 1$ $h = 1$	$g = 10$ $h = 3$	$g = 10,000$ $h = 10,000$	log-likelihood
1	0.0000	0.0000	0.0000	-4400.87
2	0.0000	0.0000	0.0000	-4305.54
3	0.0000	0.0000	0.0000	-4233.42
4	0.0000	0.0000	0.0000	-4172.52
5	0.0000	0.0000	0.0000	-4136.45
6	0.0000	0.0000	0.0000	-4111.06
7	0.0000	0.0000	0.0000	-4092.65
8	0.0000	0.0000	0.0000	-4068.25
9	0.9464	0.9533	0.9532	-4055.57
10	0.0536	0.0467	0.0468	-4058.62
contiguity	0.0000	0.0000	0.0000	-4363.82

5.3 Model comparison over alternative spatial weights

The impact of prior information on posterior inferences regarding the appropriate spatial weight matrix with a reasonably large sample of 3,107 US county observations is illustrated using the voter participation data from Pace and Barry (1997) described in the previous section. Three alternative normal priors on β and α centered on prior means of zero were employed, since the variables were log-transformed allowing interpretation of the coefficients as elasticities. A relatively uninformative prior on σ based on $\nu = \bar{s}^2 = 0.01$ was used. Three settings of a diagonal prior variance-covariance matrix $g \cdot I_k$, with $g = \{1, 10, 10^4\}$ for β were used, reflecting tight, medium and loose implementation of the prior mean of zero. A set of three corresponding variances, $h = \{1, 3, 10^4\}$ were applied to the normal prior on α, again reflecting a continuum from tight to loose.

Posterior model probabilities were calculated for models based on weight matrices with 1 to 10 nearest neighbors, and a weight matrix based on first-order spatial contiguity using all three sets of priors. These are presented in Table 5, where we see that the priors had little impact on the posterior model probabilities which all pointed to a model based on nine nearest neighbors. The last column in the table shows the log-likelihood function values indicating that a similar conclusion would arise from a likelihood-based approach to inference regarding the appropriate spatial weight structure.

The ability of any specification test to detect the true model structure depends on a host of issues such as signal/noise in the data generating process, the prior information employed, and in the case of the MESS model the value of the spatial dependence parameter α. In the case of weak spatial dependence, identification of the appropriate spatial weight structure will be difficult. As an example, a sample of 258 European Union regions was used to generate a y-vector based on a MESS model with the spatial weight matrix consisting of the 5 nearest neighbors.

A series of 10 MESS models were generated based on varying α values from -1.2 to 0.6 in 0.2 increments. Other parameters were held constant at: $\sigma^2 = 1, \beta_i = 1, i = 1, \ldots, 3$ and the matrix X was a set of random standard normal deviates. Three priors described as 'loose','medium' and 'tight' in the previous section were used for the parameters β, producing similar results. A normal prior for α with mean zero and large standard deviation

Table 6: Posterior model probabilities for alternative weights

Neighbors/α	-1.2	-1	-0.8	-0.6	-0.4	-0.2	0	0.2	0.4	0.6
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
2	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00
3	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.02	0.00	0.00
4	0.00	0.00	0.00	0.00	0.04	0.10	0.00	0.03	0.00	0.00
5$\star$	1.00	1.00	0.99	0.96	0.46	0.34	0.01	0.17	0.93	0.99
6	0.00	0.00	0.00	0.03	0.11	0.09	0.03	0.06	0.05	0.00
7	0.00	0.00	0.00	0.00	0.04	0.22	0.08	0.16	0.00	0.00
8	0.00	0.00	0.00	0.00	0.12	0.05	0.15	0.27	0.00	0.00
9	0.00	0.00	0.00	0.00	0.02	0.02	0.20	0.11	0.00	0.00
10	0.00	0.00	0.00	0.00	0.03	0.03	0.10	0.04	0.00	0.00
11	0.00	0.00	0.00	0.00	0.13	0.05	0.12	0.03	0.00	0.00
12	0.00	0.00	0.00	0.00	0.01	0.02	0.25	0.04	0.00	0.00

of 100 was employed, and the prior for σ^2 was set to the uninformative $\nu = \bar{s}^2 = 0.01$ as in the previous section.

Posterior model probabilities based on tight priors are presented in Table 6, for models associated with 1 to 12 nearest neighbor weight matrices and the alternative values of α. In the presence of relatively strong positive or negative spatial dependence (α less than -0.4 and greater than 0.2), there are high posterior probabilities associated with the correct model based on 5 nearest neighbors. When spatial dependence becomes weak, the posterior model probabilities become more uniformly distributed across models associated with alternative spatial weight matrices, indicating a difficulty in distinguishing the appropriate weight matrix structure. For the case of $\alpha = 0$, posterior model probabilities are nearly equal to the uniform prior probabilities assigned.

To illustrate the impact of sample size on this type of model comparison experiment, we present results from a large sample of 59,267 US Census tract observations. The data generating process used for the experiment was a SAR model, again based on 5 nearest neighbors. Alternative models based on 1 to 10 nearest neighbors were included in the model comparison experiment. Again, spatial dependence was varied from $\rho = -0.8$ to $\rho = 0.8$ in 0.2 increments. Because of the large sample size entirely diffuse priors on β and σ were used with a standard normal prior distribution on α.

Posterior model probabilities are presented in Table 7, where posterior model probabilities of unity appear for the 5 nearest neighbors weight matrix (the correct model) for all non-zero values of the parameter ρ. As in the previous illustration, zero spatial dependence leads to a relatively uniform set of posterior model probabilities. Intuitively, as the sample size increases there is additional sample data information with which to determine the correct spatial weight matrix specification, even in the case of very weak spatial dependence.

5.4 Model comparison for alternative explanatory variables

To illustrate the MC^3 methodology in the context of the MESS model, we utilize a dataset from Fernandez, Ley and Steel (2001a) and Sala-i-Martin (1997) concerning cross-county growth regressions. The growth rates of per capita GDP from 1960 to 1992 for a group of 72 countries and 21 explanatory variables were examined using the Zellner $g-$prior discussed

Table 7: Posterior model probabilities for a very large sample

Neighbors/ρ	-0.8	-0.6	-0.4	-0.2	0	0.2	0.4	0.6	0.8
1	0.00	0.00	0.00	0.00	0.0433	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	0.0635	0.00	0.00	0.00	0.00
3	0.00	0.00	0.00	0.00	0.0686	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.0907	0.00	0.00	0.00	0.00
5⋆	1.00	1.00	1.00	1.00	0.0919	1.00	1.00	1.00	1.00
6	0.00	0.00	0.00	0.00	0.1052	0.00	0.00	0.00	0.00
7	0.00	0.00	0.00	0.00	0.1055	0.00	0.00	0.00	0.00
8	0.00	0.00	0.00	0.00	0.1371	0.00	0.00	0.00	0.00
9	0.00	0.00	0.00	0.00	0.1545	0.00	0.00	0.00	0.00
10	0.00	0.00	0.00	0.00	0.1396	0.00	0.00	0.00	0.00

in section 4.2. The control parameter g for the $g-$prior was set to $1/k^2$, reflecting the prior suggested by Fernandez *et al.* (2001a), since $k^2 > n$ in this application.

Fernandez *et al.* (2001a) considered a set of 41 possible explanatory variables, but many of these represent categorical or other variables that proxy spatial locations. For example, there are dummy variables for Sub-Saharan countries, the absolute latitude of the country, Spanish, French, British colony dummies, and Latin American country dummies. The use of the MESS model, which explicitly captures spatial dependence, should reduce the need to consider these geographical variables. The growth rates were multiplied by 100 for ease of presentation, so the parameter estimates should be divided by 100 for comparability with other studies using this dataset.

Given 21 explanatory variables there are $2^{21} = 2,097,152$ possible ways of combining these to form alternative models. An MC^3 sampling scheme was used to produce sequential Markov Chain Monte Carlo (MCMC) draws from the complete sequence of conditional distributions for the parameters in the model. For the parameters β and σ these take the form of multivariate normal and chi-squared distributions. For the spatial dependence parameter α, univariate numerical integration was used to determine the conditional posterior and a draw was carried out using inversion.

We follow Fernandez *et al.* (2001a) and take advantage of the fact that the log-marginal posterior can be stored for all unique models encountered during MCMC sampling, allowing exact computation of posterior model probabilities. In addition, since the chain indicates which models have posterior support, the empirical frequencies of visits by the chain provide another method of determining posterior model probabilities. A high positive correlation between these two model probability measures can be used to examine convergence of the MCMC sampling scheme.[12]

For our purposes of demonstration, we carried out 50,000 draws, with 10,000 burn-in draws. This sampling run uncovered 29,716 unique models, with only 57 models exhibiting

[12] Fernandez *et al.* (2001a) report running chains of 2 million draws and finding 149,507 models visited, but 25,219 models cover 90% of the posterior model probability. Only 76 models exhibited posterior probabilities over 0.1%, and they all included between 6 and 12 regressors. They also report that smaller runs with 500,000 draws and 100,000 burn-in draws produced similar results. For example, the 76 best models were exactly the same as those based on the run of 2 million draws with 1 million burn-in draws. Their problem involving 41 explanatory variables leads to a much large model space consisting of: $2^{41} = 2,199,023,255,552$ possibilities, or 1 million times the number of possible models considered here.

Table 8: Posterior probabilities for variables inclusion in the model

	BMA Post. Prob	Frequency of visits	Frequency of appearance in the 57 highest posterior probability models
PRIMARY SCHOOL ENROLLMENT	0.3806	0.3773	3
POPULATION GROWTH	0.3961	0.3804	2
CIVIL LIBERTIES	0.4142	0.4118	2
SIZE OF LABOR FORCE	0.4144	0.4019	5
PUBLIC EDUCATION SHARE	0.4245	0.4091	2
POLITICAL RIGHTS	0.4278	0.4127	3
S.D. OF BLACK-MARKET PREMIUM	0.4383	0.4244	4
WAR DUMMY	0.4455	0.4366	6
FRACTION GDP IN MINING	0.4768	0.4653	7
HIGHER EDUCATION ENROLLMENT	0.5490	0.5239	21
REVOLUTIONS AND COUPS	0.5548	0.5276	12
NON-EQUIPMENT INVESTMENT	0.5831	0.5546	23
EXCHANGE RATE DISTORTIONS	0.5981	0.5751	18
# YEARS OPEN ECONOMY	0.7837	0.7604	53
ECONOMIC ORGANIZATION	0.7942	0.7631	55
AGE	0.9001	0.8791	57
RULE OF LAW	0.9020	0.8749	57
EQUIPMENT INVESTMENT	0.9482	0.9392	57
RATIO OF WORKERS TO POPULATION	0.9579	0.9510	57
LIFE EXPECTANCY	0.9627	0.9548	57
GDP LEVEL IN 1960	0.9728	0.9661	57

posterior probabilities over 0.1%, similar to the results in Fernandez *et al.* (2001a). Despite this relatively short sampling time, the two measures of posterior model probability based on frequency of visits and the BMA posterior calculation exhibited a correlation of 0.9993, suggesting no problems with convergence. These two measures are presented in Table 8.[13] In addition, the table presents the frequency distribution of variables appearing in the 57 highest posterior probability models.

In table 8 five variables exhibited posterior probabilities of inclusion greater than 90%, with three of these (GDP LEVEL IN 1960, LIFE EXPECTANCY AND EQUIPMENT INVESTMENT) matching the variables found by Fernandez *et al.* (2001a). Departures from those findings are the RATIO OF WORKERS TO POPULATION, which exhibited an inclusion probability less than 3 percent in Fernandez *et al.* (2001a) and the RULE OF LAW variable which had a probability of inclusion of 51.6 percent. Six variables appear in all 57 of the highest posterior probability models and two additional variables appear in 53 and 55 of these models. Four other variables appear in between 12 and 23 models, with the remaining variables appearing in 7 or less models. These results seem consistent with those in Fernandez *et al.* (2001a) who report between 6 and 12 variables in their highest posterior probability models. Due to the lack of direct comparability to the careful and extensive work of Fernandez *et al.* (2001a), it seems imprudent to make detailed comparisons of these results to theirs.

Given posterior model probabilities, one can construct a model averaged set of estimates and measures of dispersion. Posterior means and standard deviations of the parameter

[13] As in the case of Fernandez *et al.* (2001a), a second run of 50,000 draws produced an identical set of 57 high posterior probability models, as did smaller runs based on only 20,000 draws.

Table 9: Bayesian model averaged estimates

Variable	Coefficient	std deviation	Implied $t-$statistic $HO : \beta = 0$
CONSTANT	-0.0000	-0.0000	-0.04
GDP LEVEL IN 1960	-0.8873	0.1255	-7.07
PRIMARY SCHOOL ENROLLMENT	0.0021	0.0047	0.45
LIFE EXPECTANCY	0.6070	0.1328	4.57
HIGHER EDUCATION ENROLLMENT	-0.0433	0.0309	-1.40
PUBLIC EDUCATION SHARE	0.0004	0.0017	0.24
REVOLUTIONS AND COUPS	0.0177	0.0142	1.25
WAR DUMMY	-0.0050	0.0062	-0.81
POLITICAL RIGHTS	-0.0018	0.0038	-0.47
CIVIL LIBERTIES	-0.0006	0.0026	-0.23
FRACTION GDP IN MINING	0.0065	0.0074	0.88
ECONOMIC ORGANIZATION	0.1358	0.0638	2.13
EXCHANGE RATE DISTORTIONS	-0.0307	0.0216	-1.42
EQUIPMENT INVESTMENT	0.3172	0.0867	3.66
NON-EQUIPMENT INVESTMENT	0.0449	0.0295	1.52
S.D. OF BLACK-MARKET PREMIUM	-0.0035	0.0040	-0.87
# YEARS OPEN ECONOMY	-0.1191	0.0593	-2.01
AGE	-0.2108	0.0840	-2.51
RULE OF LAW	0.2845	0.0952	3.00
POPULATION GROWTH	0.0008	0.0030	0.27
RATIO OF WORKERS TO POPULATION	-0.2957	0.0601	-4.92
SIZE OF LABOR FORCE	-0.0037	0.0047	-0.78
α	-0.3324	0.0955	-3.48

estimates can be used to construct a posterior probability weighted set of estimates that incorporate model uncertainty. This was done using the 57 models exhibiting posterior probability above 0.1%, with the results presented in Table 9. In Table 9 the spatial dependence parameter α is more than three standard deviations away from zero suggesting the presence of spatial dependence in the sample data.

Only five explanatory variables exhibited parameter estimates three or more standard deviations away from zero, and these five variables exhibited inclusion probabilities greater than 90%. Estimates for three additional variables were more than two standard deviations away from zero, AGE, ECONOMIC ORGANIZATION, and # YEARS OPEN ECONOMY. These three variables exhibit inclusion probabilities between 78 and 90% and appear in 57, 55, 53 of the 57 highest posterior probability models respectively (see Table 8).

6 Conclusion

Researchers frequently rely on spatial autoregressive models in cases where the dependent variable exhibits spatial dependence. We introduce another spatial dependence model based on the matrix exponential (MESS) which replaces the geometric pattern of decay in the SAR model with one of exponential decay. This specification has theoretical as well as computational advantages over the spatial autoregressive specification. Theoretical and numeric advantages arise from the ease of inversion, differentiation, and integration of the matrix exponential. Moreover, the matrix exponential is always a positive definite matrix.

Finally, the matrix exponential has a simple matrix determinant which vanishes for the common case of a spatial weight matrix with a trace of zero. This simplification was used to produce a closed-form solution for maximum likelihood estimates, and to provide Bayesian estimates based on univariate numerical integration of a scalar polynomial expression.

We illustrated how the analytical and computational advantages of MESS can be exploited in maximum likelihood and Bayesian modelling situations. Bayesian modelling can proceed using simple numerical integration or MCMC sampling. Conventional regression diagnostics as well as Bayesian model comparison methods can be carried out by drawing on simple extensions of the existing regression model literature. For example, in the case of a normal-gamma prior for β, σ and a normal prior for the spatial dependence parameter, we demonstrate that as the priors become diffuse, the modal posterior parameters equal maximum likelihood estimates, a result that mirrors standard Bayesian linear regression.

In terms of statistical performance of the specification, simulated and applied examples demonstrated that MESS can be applied to modelling situations where conventional spatial autoregressive methods have been used. If the sample size is large or the priors are relatively uninformative, estimates and inferences similar to those from maximum likelihood estimation of conventional spatial autoregressive models arise. We demonstrated the ability of MESS in the model specification arena using illustrations involving posterior inference regarding the correct weight matrix specification as well as inference regarding explanatory variables for inclusion in the model. These demonstrations suggest that MESS holds the potential to produce simultaneous estimation and inference regarding both the spatial weights and explanatory variables. This would provide a unified and formal treatment of the most important aspects of model uncertainty for spatial regression modelling.

References

Anselin, L. (1988) *Spatial Econometrics: Methods and Models*, (Dorddrecht: Kluwer Academic Publishers).

Barry, Ronald, and R. Kelley Pace, (1999) "A Monte Carlo Estimator of the Log Determinant of Large Sparse Matrices," *Linear Algebra and its Applications*, Volume 289, Number 1-3, pp. 41-54.

Berger, J. O. and Bernardo, J.M. (1989) "Estimating a product of means: Bayesian analysis with reference priors," *Journal of the American Statistical Association* Volume 84, pp. 200-207.

Berger, J. O. and Pericchi, L.R. (1996a), "The intrinsic Bayes factor for model selection and prediction," *Journal of the American Statistical Association*, Volume 91, pp. 109-122.

Berger, J. O. and Pericchi, L.R. (1996b). "The intrinsic Bayes factor for linear models", (with discussion). In *Bayesian Statistics 5* (J. M. Bernardo et al, eds.). (London: Oxford University Press).

Berger, J. O., Pericchi, L. R. and Varshavsky, J. A. (1998) "Bayes factors and marginal distributions in invariant situations," *Sankhya, Series A, Indian Journal of Statistics*, Volume 60, pp. 307-321.

Brueckner, J.K., (2003) "Strategic interaction among governments: an overview of empirical studies," *International Regional Science Review* Volume 26, number 2, pp.

Chen, Jian-Shen, and Robert Jennrich (1996), "Signed Root Deviance Profile and Confidence Intervals in Maximum Likelihood Analysis," *Journal of the American Statistical Association*, Volume 91, p. 993-998.

Chiu, Tom Y.M., Tom Leonard, and Kam-Wah Tsui, (1996) "The Matrix-Logarithmic Covariance Model," *Journal of the American Statistical Association*, Volume 91, pp. 198-210.

Christensen, Ronald, (1996) *Plane Answers to Complex Questions*, Second Edition, (New York: Springer-Verlag).

Christensen, Ronald, Wesley Johnson, and Larry Pearson, (1992) "Prediction Diagnostics for Spatial Linear Models," *Biometrika*, Volume 79, p. 583-591.

Denison, D.G.T, Mallick B.K. and Smith, A.F.M (1998) "Automatic Bayesian curve fitting," *Journal of the Royal Statistical Society, Series B*, Volume 60, pp. 333-350.

Fan, Jianqing, Hui-Nien Hung, and Wing-Hung Wong, (2000) "Geometric Understanding of Likelihood Ratio Statistics," *Journal of the American Statistical Association*, Volume 95, pp. 836-841.

Fernández,Carmen, Eduardo Ley, and Mark F. J. Steel, (2001a) "Model uncertainty in cross-country growth regressions," *Journal of Applied Econometrics*, Volume 16, number 5, pp. 563 - 576.

Fernández,Carmen, Eduardo Ley, and Mark F. J. Steel, (2001b) "Benchmark priors for Bayesian model averaging", *Journal of Econometrics* Volume 100, number 2, pp. 381-427.

Gelfand, A. E., Dey, D. K. and Chang, H. (1992),"Model determination using predictive distributions with implementation via sampling-based methods," In *Bayesian Statistics 4* (J. M. Bernardo et al, eds.). (London: Oxford University Press).

Gentle, James, (1998) *Numerical Linear Algebra for Applications in Statistics*, (New York: Springer-Verlag).

Haining, Robert, (1994) "Diagnostics for Regression Modeling in Spatial Econometrics," *Journal of Regional Science*, Volume 34, pp. 325-341.

Hendry, David, Adrian Pagan, and Denis Sargan, (1984) "Dynamic Specification," in: Griliches, Z., and M. Intrilligator, eds. *Handbook of Econometrics, Volume 2*, (Amsterdam: North-Holland), pp. 1023-1100.

Hepple, L.W. (1995a), “Bayesian techniques in spatial and network econometrics: 1. Model Comparison and posterior odds”, *Environment and Planning A,* Volume 27, pp. 447-469.

Hepple, L.W. (1995b), “Bayesian techniques in spatial and network econometrics: 2. Computational methods and algorithms”, *Environment and Planning A,* Volume 27, pp. 615-644.

Horn, Roger, and Charles Johnson, (1993) *Matrix Analysis*, (New York: Cambridge University Press).

Kelejian, H., and I.R. Prucha, (1998) “A Generalized Spatial Two-Stage Least Squares Procedure for Estimating a Spatial Autoregressive Model with Autoregressive Disturbances”, *Journal of Real Estate and Finance Economics*, Volume 17, Number 1, pp. 99-121.

Kelejian, H., and I.R. Prucha, (1999) “A Generalized Moments Estimator for the Autoregressive Parameter in a Spatial Model”, *International Economic Review*, Volume 40, pp. 509-533.

LeSage, James P. (1997) “Bayesian Estimation of Spatial Autoregressive Models,” *International Regional Science Review*, Volume 20, number 1&2, pp. 113-129.

LeSage, James P. (2000) “Bayesian Estimation of Limited Dependent variable Spatial Autoregressive Models,” *Geographical Analysis*, Volume 32, number 1, pp. 19-35.

Lindley, D.V. (1957) “A Statistical Paradox,” *Biometrika*, Volume 44, pp. 187-192.

López-Bazo E., Vayá E., Mora A., Suriñach J. (1999), “Regional Economic Dynamics and Convergence in the European Union”. *Annals of Regional Science*, Volume 33, pp. 343-370.

Madigan, D., and J. York (1995), “Bayesian Graphical Models for Discrete Data.” *International Statistical Review*, Volume 63, pp. 215-232.

Mardia, K.V., and A.J. Watkins, (1989) “On Multimodality of the Likelihood in the Spatial Linear Model,” *Biometrika*, Volume 76, pp. 289-295.

Martin, Richard J. (1992) “Leverage, Influence, and Residuals in Regression Models when Observations are Correlated,” *Communications in Statistics: Theory and Methods*, Volume 21, pp. 1183-1212.

Ord, J.K., (1975) “Estimation Methods for Models of Spatial Interaction,” *Journal of the American Statistical Association*, Volume 70, pp. 120-126.

Pace, R. Kelley, and Ronald Barry, (1998) “Simulating Mixed Regressive Spatially Autoregressive Estimators,” *Computational Statistics*, Volume 13, Number 3, pp. 397-418.

Press, William, Saul Teukolsky, William Vetterling, and Brian Flannery, (1996) *Numerical Recipes in Fortran 77*, second edition, (New York: Cambridge University Press).

Raftery, A.E., D. Madigan, and J.A. Hoeting, (1997) "Bayesian Model Averaging for Linear Regression Models," *Journal of the American Statistical Association,* Volume 92, pp. 179-191.

Richardson, S. and Green, P.J. (1997) "On Bayesian analysis of mixtures with an unknown number of components", *Journal of the Royal Statistical Society, series B*, Volume 59, pp. 731-792.

Sala-i-Martin, X.X. (1997), "I Just Ran Four Million Regressions." Mimeo, Columbia University.

Sidje, Roger B., (1998) "Expokit: a Software Package for Computing Matrix Exponentials," *ACM Transactions on Mathematical Software*, Volume 24, pp. 130-156.

Smith, A.F.M and Roberts, G.O. (1993) "Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods," *Jounrla of the Royal Statistical Society, series B*, Volume 55, pp. 3-23.

Speigelhalter, D.J., Best, N.G., Carlin, B.P. and van der Linde, A. (2001) "Bayesian measures of model complexity and fit", *Journal of the Royal Statistical Society, Series B* (to appear).

Spiegelhalter, D. J. and Smith, A. F. M. (1982). "Bayes factors for linear and log-linear models with vague prior information," *Journal of the Royal Statistical Society Series B*, Volume 44, pp. 377-387.

Strang, Gilbert, (1976) *Linear Algebra and its Applications*, (New York: Academic Press).

Warnes, J.J., and Brian Ripley, (1987) "Problems with Likelihood Estimation of Covariance Functions of Spatial Gaussian Processes," *Biometrika*, Volume 74, pp. 640-642.

Zellner, Arnold. (1971) *An Introduction to Bayesian Inference in Econometrics*. (New York: John Wiley & Sons.)

Zellner, A. (1986) "On assessing prior distributions and Bayesian regression analysis with g-prior distributions," *In Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti* (North-Holland/Elsevier: Amsterdam).

Appendix A: Uniqueness of the maximum likelihood solution

To narrow the possible number of solutions, we turn to the second order conditions. Positive definite Q would usually prove sufficient for an interior solution, but the vector v embodies a polynomial constraint. Therefore, we elaborate on the second order conditions taking into account the constraints imposed by the structure of v. Consider the second derivative of the sum-of-squared errors with respect to α shown in (43).

$$\frac{d^2 v'Qv}{d\alpha^2} = \sum_{i=3}^{2q-1} c_i(i-1)(i-2)\alpha^{i-3} = 2\left[(\frac{dv}{d\alpha})'Q(\frac{dv}{d\alpha}) + v'Q(\frac{d^2v}{d\alpha^2})\right] \tag{43}$$

The first term inside the brackets is positive because it represents a positive definite quadratic form. We can rewrite the second term in brackets as shown in (44),

$$v'Q\left(\frac{d^2v}{d\alpha^2}\right) = v'(QA)v \tag{44}$$

where A equals,

$$A = \left(\frac{1}{\alpha^2}\right)\begin{pmatrix} 0 & & & & \\ & 0 & & & \\ & & (i-1)(i-2) & & \\ & & & \ddots & \\ & & & & (q-1)(q-2) \end{pmatrix}. \tag{45}$$

The minimum value of α depends on the eigenvalues of $v'(QA)v$. Note that Q is positive definite and the real diagonal (and thus Hermitian) matrix A in (45) has two zero and $(q-2)$ positive eigenvalues. Horn and Johnson (1993, p. 465) state in Theorem 7.6.3 that the product of a positive definite matrix Q and a Hermitian matrix A has the same number of zero, positive, and negative eigenvalues as A. Hence, QA must have two zero and $(q-2)$ positive eigenvalues. Therefore, $v'(QA)v$ is positive semi-definite implying that $v'QAv$ has a minimum value of 0. Since the first term in brackets in (18) always has a positive value, the entire expression in (1) has a positive value and thus $v'Qv$ is positive definite and strictly convex in α. Hence, if an interior solution exists to the first order conditions, it must be unique.

There exists an interior solution to the first order conditions. To see this, examine the highest degree term in $P(\alpha)$, from (16) shown in (46).

$$T_{\max} = \frac{\alpha^{2(q-1)}[y'D'^{(q-1)}MD^{(q-1)}y]}{(q!)^2} \tag{46}$$

The term in brackets is the contribution to the overall sum-of-squared errors from the last term in the truncated Taylor's series and must be positive. Since $\alpha^{2(q-1)}$ is even in α, only the magnitude and not the sign of α matter for this result. Since $\lim_{|\alpha|\to\infty} T_{\max} \to \infty$, implies $\lim_{|\alpha|\to\infty} v'Qv \to \infty$, there exists an interior solution to the first order conditions.

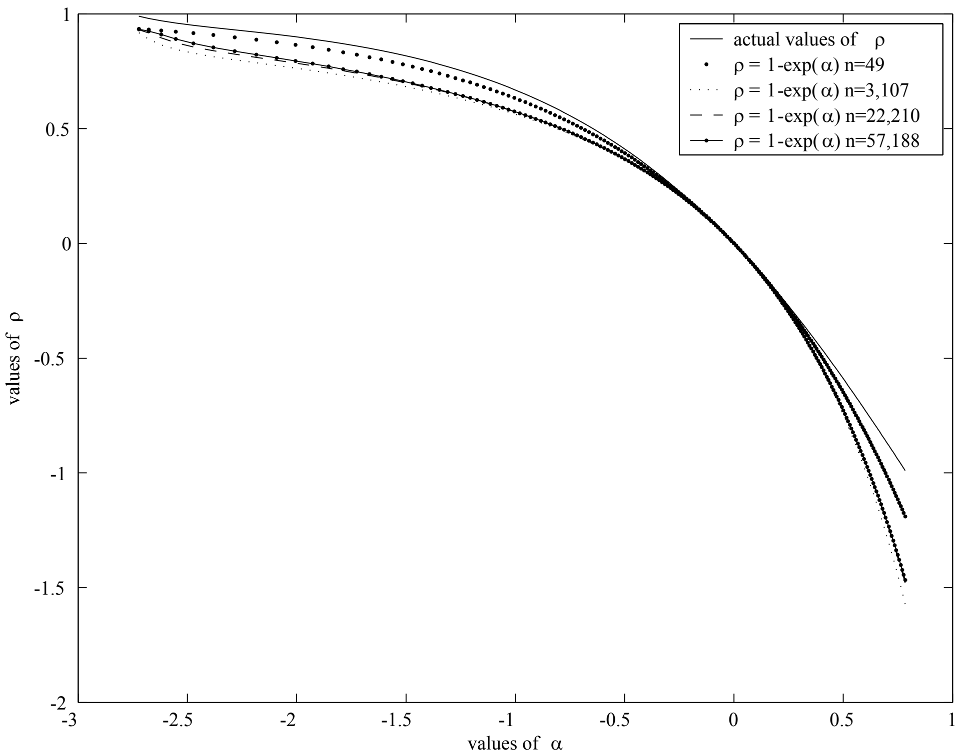

Figure 1: Correspondence between α and ρ

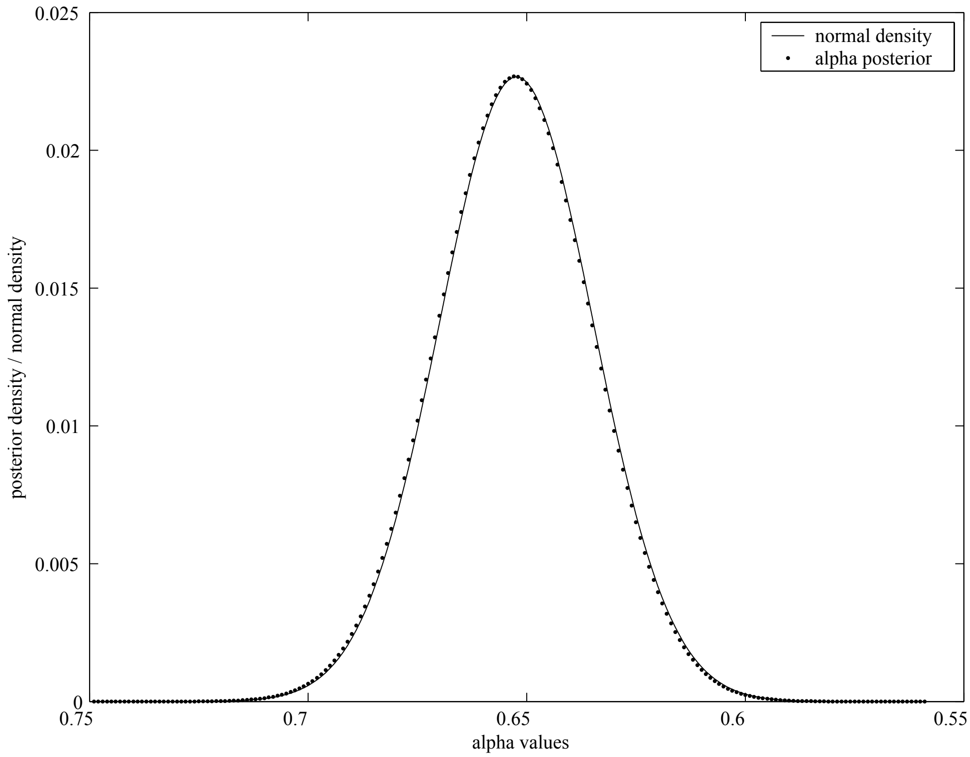

Figure 2: Correspondence normal pdf and posterior pdf for α