Bayesian MCMC Inference for Spatial Models

Jeffrey A. Mills
Department of Economics
University of Cincinnati
Cincinnati, Ohio 45221
jeffrey.mills@uc.edu
and
Olivier Parent
Department of Economics
University of Cincinnati
Cincinnati, OH 45221
olivier.parent@uc.edu

February 16, 2012

Abstract

This paper provides a survey of the recent literature on Bayesian inference methods in regional science. This discussion is presented in the context of the Spatial Durbin Model (SDM) with heteroskedasticity as a canonical example. Overall performance of different hierarchical models is analyzed. We extent the benchmark specification to the dynamic panel data model with spatial dependence. An empirical illustration of the flexibility of the Bayesian approach is provided through the analysis of the role of knowledge production and spatiotemporal spillover effects using a space-time panel data set covering 49 US states over the period 1994-2005.

1 Introduction

Applied work in regional science is increasingly confronted with the task of analyzing data that are geographically referenced and temporally correlated, with many potential predictors. Up until the 1990s, virtually all of the empirical work in regional science employed frequentist statistical methods. The landmark work by Anselin (1988) reviews this literature and provides arguably the most comprehensive coverage of spatial econometrics in regional science.

In the early 1990s, the development of Markov Chain Monte Carlo (MCMC) methods revolutionized applications of the Bayesian approach to statistical inference. The revival of interest in the Bayesian approach has rapidly extended into spatial econometrics and geo-statistics. MCMC techniques, applied creatively, allow for the sophisticated modeling of large data sets with time dependence and cross-sectional correlation. Recent developments in Bayesian methods allow fully Bayesian analyses of sophisticated multilevel models for complex geographically referenced data (Banerjee et al. 2004; LeSage and Pace 2009). This approach also offers full inference for non-Gaussian spatial data, spatiotemporal data, and, for the first time, solutions to problems of interpretation for models incorporating geographic and temporal dependence.

Analysing a variety of panel data models, Chib (2008) underlines how the approach allows for the complex analysis of continuous, censored, count and multinomial responses under weaker assumptions than required by previously developed methods. For instance, the Bayesian approach does not require the strict exogeneity assumption in the presence of endogenous covariates. Based on this panel setting, a growing number of studies examine spatial and temporal effects in multinomial or multivariate discrete response data. For example, Wang and Kockelman (2011) develop a dynamic spatial ordered probit model and use it to analyze land development intensities. Discrete choice modeling with spatial dependence has been deeply analyzed using mainly the Bayesian approach (an extensive review can be found in LeSage and Pace, 2009).

The development of new theoretical and empirical models in regional science to analyze, among other things, regional economic growth (Ertur and Koch 2007; LeSage and Fischer 2008), land use and conservation (Wang et al. 2011), industrial localization (Kakamu et al. 2011), geography of innovation (Autant-Bernard and LeSage 2011; Parent and LeSage 2008), highlights the flexibility of the Bayesian approach. There are additional problems that arise in the modeling process, such as model comparison and predictive performance, that have proven problematic in the past, but can now also be addressed in a relatively straightforward manner using Bayesian inference and MCMC methods.

The rapid growth in availability of software incorporating MCMC methods has contributed to the dissemination and use of Bayesian methods in empirical work in regional science. A wide range of toolboxes contain all the standard procedures for empirical analysis. A comprehensive collection of routines can be found in one of the best known toolboxes for spatial data analysis, the spatial econometric toolbox of James LeSage[1]. These routines are implemented within the Matlab environment and contain the most advanced tools for spatial analysis and model interpretation. An increasingly attractive alternative is based

[1] http://spatial-econometrics.com/

on the development of statistical packages in the open source R environment[2]. An extensive collection of geo-statistics toolboxes are developed using Bayesian techniques. It is also worth mentioning the significant impact of open source software such as Winbugs. This has been used to make some significant contributions to empirical analysis in regional science.

This chapter presents recent econometric advances in the treatment of complex spatial and spatiotemporal data sets, and outlines a comprehensive approach to dealing with spatial and time effects from a Bayesian econometric perspective. The main objective is to illustrate how Bayesian techniques can help to understand a number of spatial theories and empirical models that have been developed for the practice of regional science and policy analysis. This discussion is presented in the context of the Spatial Durbin Model (SDM) as a canonical example.

The SDM is presented in the next section. Because the Bayesian method is inextricably tied to MCMC sampling, we provide a brief overview of MCMC methods in sections 3, 4 and 5. Section 6 then applies MCMC methods to the SDM to demonstrate some recent Bayesian research relevant for spatial econometric modeling, particularly with regard to problems of heteroskedasticity and spatial dependence in a panel data setting. The model is extended to include time dependence and a substantive application of the methodology to regional growth models with interregional technological dependence is then provided in section 7. Lastly, section 8 summarizes and provides some concluding thoughts that relate to the future of Bayesian econometrics in regional science.

2 Spatial Regression And Prior Modeling

The Bayesian approach to spatial modeling relies extensively on the idea of a hierarchical prior which is used to model spatial dependence and heterogeneity. Suppose we have a cross-sectional sample of N independent observations y_i, $i = 1, \ldots, N$ that are linearly related to a set of $N \times k$ explanatory variables X and are believed to be spatially correlated. As a benchmark, we will start with the Spatial Durbin Model, which can be motivated by concern over omitted variables or spatial heterogeneity (see LeSage and Pace 2009). This specification includes spatial lags of the explanatory variables as well as the dependent variable.

A representation of the Bayesian SDM model is shown in (1),

$$\begin{aligned} y &= \rho W y + \iota_n \alpha + X\beta + WX\gamma + \epsilon \\ \epsilon &\sim N(0, \sigma_\epsilon^2 \Lambda^{-1}) \\ \Lambda &\equiv diag(1/\lambda_1, \ldots, 1/\lambda_n) \\ \lambda_i &\sim \chi_\nu^2/\nu, \end{aligned} \tag{1}$$

where W is a known $N \times N$ spatial weight matrix whose diagonal elements are zero, ι_n is a $1 \times N$ column vector of ones, and the strength of the spatial dependence is measured by the parameter ρ. The W matrix defines the structure of the dependence between (spatial) observational units. We also assume that W is normalized from a symmetric matrix, so that

[2] http://r-project.org/

all eigenvalues are real and less than or equal to one. Different normalization methods can be used. For example, unlike the traditional row-normalization, the spectral-normalized matrix preserves the symmetry by dividing each element by the modulus of the largest eigenvalues (Barry and Pace 1999).

We add a normal-inverse gamma prior for β and σ_ϵ, and we introduce a uniform prior distribution for the parameter ρ. Intuitively, if we were to simply treat Λ as N unrestricted parameters, a degrees of freedom problem would arise. Geweke (1993) proposes a set of N independent, identically distributed, Chi-Square distributions as prior information for the variance scalars λ_i,

$$p(\Lambda) = \prod_{i=1}^{n} Ga(\lambda_i | \frac{\nu}{2}, \frac{\nu}{2}). \tag{2}$$

The parameter ν represents the single parameter of the Gamma distribution equivalent to a Chi-Square distribution, allowing us to estimate the N variance scaling parameters λ_i by adding only a single parameter to the model. Geweke (1993) shows that this approach to modeling the disturbances is equivalent to a model that assumes a Student-t distribution for the errors. Another way to view this is that using a t distribution to deal with heteroskedasticity is equivalent to a scale mixture of normals when the mixing distribution is a Gamma distribution. That is, assuming that λ_i are independent $N(0, \sigma_\epsilon^2 \lambda_i^{-1})$ with prior for λ_i given in (2), is equivalent to the assumption that the error distribution is a weighted average of different normal distributions, each with a different variance. Additional flexibility in modeling heterogeneity can be achieved by introducing a prior hyperparameter for ν that follows an exponential distribution governing the degrees of freedom that controls thickness of the tails in the Student-t error distribution (Geweke 1993).

3 Bayesian Inference via MCMC

As can be seen from equation (1), spatial models tend to have fairly high parameter dimensionality. This is because the minimal level of complexity needed to adequately deal with variations in neighboring structure is rather high. As a result, analytical derivation of closed form expressions for Bayesian posterior distributions is not usually possible for these models. Fortunately, MCMC methods are a tailor-made solution to this problem as they provide approximations to posterior distributions in complex settings up to an arbitrary degree of numerical accuracy.

MCMC techniques allow simulation of a sample from any distribution by embedding it as a limiting distribution of a Markov chain, then simulating from the chain until it approaches equilibrium. This is essentially achieved by reverse engineering with the goal of finding a Markov chain algorithm that will ultimately converge upon the target distribution. Analogs of the law of large numbers and central limit theorems (see section 3.2 below) exist for Markov chains that ensure that most of the simulated values from a chain can be used to provide information about the distribution of interest. The degree of accuracy can then be increased arbitrarily simply by increasing the simulated sample size.

A large theoretical literature now exists that sets out the conditions under which the MCMC chain converges to the target posterior. These conditions are surprisingly weak,

though there is usually no way to guarantee that they hold in practice. However, a high degree of confidence that the Markov chain has converged can often be achieved, especially if care is taken to follow the suggestions in Geyer (2011) and employ the diagnostic tools discussed in section 5 below.

In the last two decades, powerful MCMC techniques have been developed to obtain random draws from a very wide class of conditional distributions under remarkably general conditions. Even when the conditional distributions are too complex for Gibbs MCMC, Metropolis-Hastings (MH) algorithms can be employed to ensure that the appropriate limiting distribution is maintained by rejecting unwanted moves in a chain. We will assume the availability of algorithms to draw psuedo-random numbers from a variety of standard distributions. Methods for doing so have been thoroughly studied and are now widely available in most statistical software (see Gamerman and Lopes 2006, chapter 1, for a nice exposition).

3.1 A Brief Review of MCMC Theory

Monte Carlo methods originate from early work by Stanislaw Ulum and were used during World War II at Los Alamos in the development of the atomic bomb (Metropolis and Ulum 1949). Metropolis el a. (1953) was the pioneering paper on MCMC, but it was overlooked by statisticians, partly because it was published in a chemistry journal, and partly because of the primitive level of computer technology available at the time, making computational methods prohibitively expensive for most statistical applications. Hastings (1970) generalized the Metropolis algorithm, but it was not until the late 1980s and early 1990s that widespread recognition of the practical importance of these algorithms occurred among statisticians. Geman and Geman (1984) developed the Gibbs sampler for use in image processing, and Tanner and Wong (1987) developed the data augmentation approach and were arguably the first to recognize the potential for Bayesian MCMC inference. However, it was the classic expository paper by Gelfand and Smith (1990) that brought the Gibbs sampler to the attention of a wider audience. This led to the rapid development of a generic set of MCMC tools for Bayesian inference and subsequently revolutionized the field of statistics.

Most of the theoretical developments in MCMC were achieved in the 1990s. The research drive in MCMC methods over the last decade has shifted to developing more efficient computational tools. While advances in computer technology have continued to rapidly reduce the computational costs of simulation techniques, this has led to the analysis of more and more complex models. Researchers in MCMC methods continue to push the frontier of what is currently computationally feasible and there is need for a high level of computational efficiency in this environment. Many of the spatial models employed in empirical research however, are not of such a high order of complexity, and can now be analyzed quickly and easily on any standard computer. Further, the level of complexity of the models that can be analyzed without being overly concerned with efficiency has grown dramatically in the last decade. In short, the last couple of decades have led to revolutionary changes in our ability to statistically analyze complex spatial models and problems.

3.2 Stationary Distributions and a Central Limit Theorem for MCMC

The goal of Bayesian computation is to obtain a sample of draws $\theta^{(t)}$, $t = 1, \ldots, M$, from the posterior distribution of the unknown quantity θ, with a large enough sample that

quantities of interest can be estimated with reasonable accuracy. MCMC simulation is a general method based on drawing values of θ from distributions that result in a sample from the target posterior distribution, $p(\theta|y)$. The sample is drawn sequentially, with the tth draw, $\theta^{(t)}$, depending only on the previous draw, $\theta^{(t-1)}$. This dependence on only the previous draw is the defining property of a Markov chain, so that MCMC is a practical application of Markov chain theory. Some understanding of the theory of Markov chains is thus helpful in practice, particularly in evaluating the performance and convergence of MCMC chains. This section provides a very brief review. Fuller treatments can be found in many texts, including Gelman et al. 2004; Gamerman and Lopes 2006.

A key requirement for the application of MCMC methods is the convergence of the chain to a stationary distribution. A distribution p is said to be a stationary distribution of a chain with transition probabilities $\pi = \pi(x, y)$ if $p = p\pi$. If the stationary distribution p exists and $lim_{n\to\infty} p\pi^n = p$, then, independently of the initial distribution of the chain, p^n will approach p, as $n \to \infty$.

Ergodicity concerns ensuring that the chain will visit all possible values under the support of the distribution of interest (the stationary distribution) with nonzero probability. A chain is ergodic if it is aperiodic (so it cannot get stuck cycling in one subregion of the parameter space), and positive recurrent (which essentially means that as $n \to \infty$, the probability of visiting every possible state is nonzero). For a Markov chain, $\theta = (\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(n)})$, the ergodic average of a real-valued function of θ, $h(\theta)$ is the average $\overline{h}_n = (1/n)\sum_{t=1}^{n} h(\theta^{(t)})$.

If the chain is ergodic and $E_p[h(\theta)] < \infty$ for the unique limiting distribution p, then

$$\overline{h}_n \xrightarrow{a.s.} E_p[h(\theta)] \text{ as } n \to \infty. \tag{3}$$

This result is a Markov chain equivalent of the Law of Large Numbers (see Gamerman and Lopes 2006, p.125).

If a chain is uniformly (geometrically) ergodic and $h^2(\theta)(h^{2+\epsilon}(\theta))$ is integrable with respect to p for some $\epsilon > 0$, then we can obtain a Central Limit Theorem for Markov chains,

$$\sqrt{n}\frac{\overline{h}_n - E_p[h(\theta)]}{\tau} = \sqrt{n_{\text{eff}}}\frac{\overline{h}_n - E_p[h(\theta)]}{\sigma} \xrightarrow{d} N(0,1) \text{ as } n \to \infty, \tag{4}$$

where $\sigma^2 = \text{var}(h(\theta))$ is the variance of the limiting distribution p, $\tau^2 = \sigma^2(1 + 2\sum_{k=1}^{\infty}\rho_k)$ is the limiting sample variance of the estimate $\overline{h}_n$, and

$$n_{\text{eff}} = n/(1 + 2\sum_{k=1}^{\infty}\rho_k) \tag{5}$$

is the inefficiency factor due to autocorrelation in the Markov chain, indicated by $\rho_k = \text{cov}(h(\theta^{(t)}, h(\theta^{(t-k)})/\sigma^2$. The inefficiency factor n_{eff} is used in practice to measure the 'effective' random iid sample size of the MCMC chain by replaced the theoretical autocorrelations, ρ_k, with consistent sample estimates.

Equation (3) provides theoretical support for evaluating ergodic averages as estimates, and equation (4) supports evaluating approximate confidence intervals. Tierney (1994) provides proofs of ergodicity for the Markov chains in common use for MCMC simulation, so that the above results apply. See Gamerman and Lopes (2006) for further discussion.

One further point worth highlighting is the concept of a reversible Markov chain. A chain is said to be reversible if

$$p(x)\pi(x,y) = p(y)\pi(y,x) \text{ for all } x, y \in S, \tag{6}$$

where the state space S is the appropriate subset of $\mathbb{R}^n$ representing the support of x, y.

Equation (6) is known as the 'detailed balance equation' because it equates the rates of moves through states (so balanced) for every possible pair of states (hence detailed). This leads to the key result. If there is a distribution p satisfying the detailed balance equation, (6), for an irreducible chain, then the chain is positive recurrent and reversible with stationary distribution p. Metropolis et al. (1953) showed that it is then *always* possible to construct a Markov chain with stationary distribution p by finding transition probabilities $\pi(x, y)$ satisfying (6). This provides an algorithm for constructing Markov chains that has weak requirements and so has wide applicability. The above results, in particular, convergence to the limiting distribution, the ergodic theorem and the central limit theorem, all hold for continuous state spaces with only minor technical modifications required (see Gamerman and Lopes, 2006).

The above theory provides the means by which sampling from virtually any posterior distribution p can be achieved. The basic Metropolis algorithm is to set p as the limiting distribution of an ergodic Markov chain with transition kernel π. The various algorithms that build on this, in particular Gibbs sampling and Metropolis-Hastings (MH), are concerned with various methods of providing proposal distributions π to be sampled from.

4 MCMC algorithms

The main workhorse MCMC method is Gibbs sampling, which is a special case of the MH algorithm that is very simple to use in practice. The Gibbs sampler requires knowledge of the full conditional distributions (up to an unknown constant) and so is not always usable, but simplifies the task and speeds up MCMC computations when it can be used. The MH algorithm does not require knowledge of the full conditionals and is often used in conjuction with the Gibbs sampler to obtain draws for the unknown parameters for which the full conditionals are not available.

4.1 Gibbs Sampling

The Gibbs sampler is an MCMC method that has wide applicability in spatial econometric modeling. Suppose we have a set of k parameter vectors, $\theta_1, \theta_2, \ldots, \theta_k$, where each θ_i could be a scalar or a vector of parameters (to be drawn as a block). For example, in a linear regression model $y = X\beta + e$, $e \sim N(0, \sigma^2 I)$, it is convenient to separate the unknown parameters into two blocks, treating the regression coefficients as one $(1 \times k)$ vector, so $\theta_1 = \beta$, and the variance separately as a scalar, $\theta_2 = \sigma^2$.

The Gibbs sampler can be used if we can sample from the full conditionals. The generic Gibbs sampler algorithm is to draw one value for each θ_i from its conditional distribution, and cycle through these conditionals repeatedly. For each iteration, $t = 1, 2, \ldots, M$, and arbitrary starting values $\theta_2^{(0)}, \theta_3^{(0)}, \ldots, \theta_k^{(0)}$, the algorithm is,

- draw $\theta_1^{(t)}$ from $p(\theta_1^{(t)}|\theta_2^{(t-1)}, \theta_3^{(t-1)}, \ldots, \theta_k^{(t-1)}, y)$,
- draw $\theta_2^{(t)}$ from $p(\theta_2^{(t)}|\theta_1^{(t)}, \theta_3^{(t-1)}, \ldots, \theta_k^{(t-1)}, y)$,

$\vdots$

- draw $\theta_k^{(t)}$ from $p(\theta_k^{(t)}|\theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_{k-1}^{(t)}, y)$.

The above conditional distributions are the transition distributions of a Markov chain that converges (under very general conditions) to a unique stationary target distribution that is the posterior distribution $p(\theta_i|y)$. In the linear regression example, we typically specify a normal prior distribution for β and an inverted-gamma prior for σ^2. The Gibbs sampler for this model is then to cycle through the two conditionals, drawing $\beta^{(t)}$ from $N(\beta^{(t)}|\sigma^{2\,(t-1)}, y)$ and $\sigma^{2\,(t)}$ from $IG(\sigma^{2\,(t)}, |\beta^{(t)}, y)$.

The Gibbs sample for each parameter, $\theta_i^{(t)}$, $t = 1, 2, \ldots, M$ then approximates a sample from the marginal posterior $p(\theta_i|y)$. This approximation can be made arbitrarily accurate by increasing the sample size, M. Given that it is now computationally inexpensive to obtain tens of thousands of draws on any standard computer for all but the most complex and highly dimensional models, Gibbs sampling is an easy way to draw posterior inferences concerning any unknown quantities in a model.

4.2 Metropolis-Hastings (MH)

MH algorithms are a general family of MCMC methods that use simulations from almost any arbitrary density π to actually generate draws from an equally arbitrary given target density p. Further, these algorithms allow for the dependence of π on the previous simulation, so the choice of π does not require a particularly elaborate construction *a priori*, but can take advantage of the local characteristics of the stationary distribution.

The use of a chain produced by an MCMC algorithm with stationary distribution p is fundamentally identical to the use of an *iid* sample from p in the sense that the ergodic theorem guarantees the (almost sure) convergence of the empirical average to the posterior expectation,

$$\frac{1}{M}\sum_{t=1}^{M} h(\theta^{(t)}|y) \underset{a.s.}{\longrightarrow} E_p[h(\theta|y)]$$

A sequence $\theta^{(t)}$ produced by an MCMC algorithm can thus be employed just as an iid sample. An excellent introduction to Metropolis-Hastings algorithms is provided by Chib and Greenberg (1995).

4.2.1 The Metropolis algorithm

The Metropolis (1953) algorithm is a special case of the MH algorithm which draws from a transition distribution $\pi(\theta^{(t)}|\theta^{(t-1)})$ that must be symmetric, i.e. $\pi(\theta^{(t)}|\theta^{(t-1)}) = \pi(\theta^{(t-1)}|\theta^{(t)})$. This simplifies the algorithm in that the proposed transition distribution does not need to be evaluated at each accept-reject step since it does not appear in α (see below). Starting values, $\theta^{(0)}$, are often simply arbitrarily chosen to represent a draw from a preliminary crude

approximate estimate of the posterior distribution, or are drawn from the prior distribution. Several runs of the algorithm using different starting values can be employed to diagnose convergence to the target posterior. Given starting values, for $t = 1, 2, \ldots, M$, the algorithm is,

- draw $\theta^{(t)}$ from the transition distribution $\pi(\theta^{(t)}|\theta^{(t-1)})$,
- calculate
$$\alpha = \frac{p(\theta^{(t)}|y)}{p(\theta^{(t-1)}|y)},$$
- accept $\theta^{(t)}$ with probability = $\min(\alpha, 1)$, otherwise set $\theta^{(t)} = \theta^{(t-1)}$ (i.e. keep the previous draw).

This last step is accomplished by drawing a uniform random variate r in the $[0, 1]$ interval and accepting $\theta^{(t)}$ if $\min(\alpha, 1) \geq r$.

The algorithm requires the ability to calculate the acceptance-rejection ratio α for all $(\theta^{(t)}, \theta^{(t-1)})$, and to draw $\theta^{(t)}$ from the proposal distribution $\pi(\theta^{(t)}|\theta^{(t-1)})$ for all θ and t. To prove that the sequence $\theta^{(t)}$, $t = 1, 2, \ldots$ converges to a sample from the target distribution we need, a) that the simulated sequence is a Markov chain with a unique stationary distribution, and b) that this stationary distribution equals the target posterior distribution. This holds if the Markov chain is irreducible, aperiodic and nontransient. Except for trivial exceptions, the distribution is aperiodic and nontransient for a random walk on any proper distribution, and is irreducible if the random walk has a positive probability of eventually reaching any state from any other state (i.e. the transition distribution must be able to eventually visit all possible states with nonzero probability). The acceptance step and definition of α ensures, by construction, that the stationary distribution is the target posterior (see Gelman et al., 2004).

4.2.2 Metropolis-Hasting algorithm

Hastings (1970) developed the MH algorithm as a generalization of the Metropolis algorithm such that the transition distribution is not required to be symmetric. In this case, the acceptance rule becomes,

$$\alpha = \frac{p(\theta^{(t)}|y)/\pi(\theta^{(t)}|\theta^{(t-1)})}{p(\theta^{(t-1)}|y)/\pi(\theta^{(t-1)}|\theta^{(t)})}.$$

Allowing asymmetric accept-reject rules can be useful in increasing the speed of convergence of the Markov chain. Proof of convergence to a unique stationary distribution is the same as for the Metropolis algorithm. That this stationary distribution is the target distribution follows from the definition of α. (see Gelman et al., 2004). The Gibbs sampler can also be shown to be a special case of the MH algorithm with $\alpha = 1$ always, where the transition distribution is selected to be the conditional distribution $p(\theta^{(t)}|\theta^{(t-1)}|y)$.

4.3 Choice of proposal distribution $\pi(\theta^{(t)}|\theta^{(t-1)})$

A good transition distribution is one for which, for any θ, it is easy to sample from $\pi(\theta^{(t)}|\theta^{(t-1)})$, it is easy to compute α, each accepted iteration moves a reasonable distance in the parameter space (so that the Markov chain does not move too slowly), and the rejection rate is not too high (so that the chain does not remain in the same place too often). Note that only the ratios $\pi(\theta^{(t)}|\theta^{(t-1)})/\pi(\theta^{(t-1)}|\theta^{(t)})$ and $p(\theta^{(t)}|y)/p(\theta^{(t-1)}|y)$ are required, so we only need the kernels of these distributions.

While there are an infinite variety of possibilities, there are two main methods typically used for selecting the transition distribution. Random walk MH employs a transition distribution centered at the previous draw, so the draws follow a random walk over the support of the posterior. It is the most commonly used method because of its simplicity, its validity in most situations, and it does not require in-depth *a priori* knowledge of the transition distribution. The main alternative is the independent draw MH, which can be considerably more efficient, but requires a transition distribution that is a close approximation to the target distribution. The MH acceptance step is used to correct the approximation in the independent draw MH, with the goal being to accept as many draws as possible. If the posterior can be approximated fairly accurately with some confidence, then using the independent MH makes a lot of sense. Otherwise, the random walk MH tends to be the default choice.

The random walk MH with a normal transition kernel centered on the current draw, and with covariance matrix $= c^2\hat{\Sigma}$, where $\hat{\Sigma}$ is an approximate estimate of the posterior covariance matrix, has transition matrix

$$\pi(\theta^{(t)}|\theta^{(t-1)}) \sim N(\theta^{(t-1)}, c^2\hat{\Sigma}).$$

The algorithm is then,

- Start with $\theta^{(0)}$
- Draw $\theta^{(t)} = \theta^{(t-1)} + \varepsilon$, $\varepsilon \sim N(0, c^2\hat{\Sigma})$
- Compute $\alpha = \min\{1, p(\theta^{(t)}|y)/p(\theta^{(t-1)}|y)\}$
- With probability α, accept $\theta^{(t)}$, otherwise set $\theta^{(t)} = \theta^{(t-1)}$
- Repeat as necessary

The most efficient choice of the scale term for the normal random walk MH is $c \approx 2.4/\sqrt{k}$, where k is the dimension of θ (the number of parameters). This parameter, c, can be tuned by initial runs of the MH algorithm so that the acceptance rate is between 0.2 and 0.5, with the upper end appropriate in one dimension and the lower end for higher dimensions ($k > 5$), according to Gelman, et al. (2004). While this algorithm can be improved in many ways, it has proved effective in many problems even with moderately large $k \lesssim 50$.

The independent draw MH takes the transition distribution to be independent of the current chain, so $\pi(\theta^{(t)}|\theta^{(t-1)}) = \pi(\theta^{(t)})$, and $\theta^{(t)}$ is drawn directly from this distribution, replacing the random walk step in the above algorithm. If $\pi(\theta^{(t)})$ is a good approximation to $p(\theta^{(t)}|y)$, then most draws will be accepted and we obtain a chain with almost no autocorrelation.

5 Practical considerations

In practical application, both MCMC and Bayesian inference involve a number of choices concerning various parameters that must be selected *a priori*. The need to select prior distributions has, at least in the past, been a conceptual hurdle that slowed the widespread acceptance of Bayesian theory. With regard to MCMC, choice of burn-in sample size, tuning acceptance-rejection rate, length of MCMC chain needed, whether to use one chain or parallel chains, use every subsequent (accepted) draw or only keep every kth draw (and hence choose k), and appropriate choice and monitoring of convergence diagnostics, represent only a partial list of the decisions faced by the applied researcher.

Fortunately, most of the anguish over these questions that was present in the 1990s has subsided as a combination of theoretical advances and practical experience provided reasonable answers. Bayesian inference and MCMC techniques have something of a parallel recent history in this regard. Development and extensive use of a widely accepted standard menu of relatively noninformative proper priors, coupled with demonstration of the robustness of posterior inference to reasonable variations in the parameters of these priors, along with many practical examples of their use, has essentially eliminated the controversy over the use of priors and hence Bayesian inference (see, for example, Gelman et al., 2004). During the same period, appropriate procedures and choices for setting up, fine tuning and monitoring MCMC chains have become routine.

The two main practical issues that arise when using MCMC are:

1. The early iterations can be misrepresentative of the target distribution since approximate convergence is likely to not have been reached yet, so inclusion of these early iterations will influence the posterior inference. We must therefore be sure to run the simulation algorithm for long enough to be confident that approximate convergence has been achieved and discard the early (burn-in) portion of the sample.

2. The Markov chain can often be correlated. Inference from correlated draws is less precise than from the same number of independent draws because there is less new information in each correlated draw. Correlation in the draws can therefore make the sampling algorithm inefficient if a large number of draws is necessary to achieve a relatively small effective equivalent sample size of independent draws. To monitor this, we view the autocorrelation function (ACF) and calculate the effective sample size, (5).

We outline these procedures and give further references below. Geyer (2011) is an essential reference for anyone using MCMC methods in practice.

5.1 Setting up and monitoring MCMC chains

Theoretically at least, many of the apparent problems that were of concern initially have turned out to be easily resolved. There is no theoretical justification for using any burn-in period, using parallel chains instead of just one chain, not using all subsequent draws or even for many of the convergence diagnostics originally developed. The short answer to all these issues is that one should simply run one chain for a longer time (number of iterations) to gain more confidence concerning convergence. Geyer (2011) argues that using a single longer chain is the best approach once variations in starting values have been explored. If long burn-in periods are required, or if the chains have very high autocorrelations, using

a number of smaller chains may result in each not being long enough to be of any value. Where nonconvergence could be an issue (i.e. nonstandard problems), Geyer recommends at least one run of an MCMC chain overnight; "what better way for your computer to spend its time?" (Geyer 2011, p.19).

The Gibbs sampler is the simplest of the MCMC algorithms and so is usually employed if sampling from the conditional posterior distributions is possible. If it is not possible to use the Gibbs sampler, the random walk Metropolis algorithm provides a relatively simple way to obtain an MCMC sample since we do not need to evaluate the transition distribution in the acceptance step. The computational power now available to the average user is such that obtaining MCMC sample sizes up to order 10^6 is already a fairly trivial task for many standard models. As a result, efficiency is no longer a real concern in many practical applications. In addition, a few easily implemented diagnostic tools have become standard, mainly:

(a) visual inspection of the chain itself (a simple time plot) to observe if the chain appears to have settled into a stationary path,

(b) inspection of the ACF for the chain to check for excessive time dependence, requiring a larger number of draws (checking the effective sample size of independent draws by viewing the ACF for every kth draw),

(c) initially running the chain several times from a diverse set of starting values to check if the chain converges to the same stationary path each time,

(d) tuning the acceptance rate for any MH steps to be somewhere between about 0.2 to 0.5, and

(e) calculation of numerical standard errors (n.s.e.) and an estimate of the effective sample size, n_{eff}, from (5).

A number of excellent monographs now exist that cover these issues in far more detail than is possible here. Of particular relevance for spatial modeling are Chib (2008) and especially LeSage and Pace (2009).

5.2 Other tools and post sampling inference

When running an MCMC chain, the number of iterations should never be fixed in advance. Deciding on the length of an MCMC run is a sequential process where the MCMC chains are examined after pilot runs and new simulations (or new samplers) are chosen on the basis of these pilot runs. For many situations, an MCMC sample of 100 independent draws is sufficient for reasonable posterior summaries, so even with a fairly high degree of correlation in the chain, several thousand draws are generally more than sufficient for accurate posterior inference, provided we are confident that the chain has converged (see Gelman et al. 2004). Further, we can compare sample standard errors with numerical standard errors to ensure the numerical accuracy is adequate, and run the chain for longer if it is not.

Once an MCMC sample is obtained, standard sample estimates of posterior moments and quantiles can be calculated for the unknown quantities directly, e.g. the posterior mean

of any function, $h(\theta)$, of the unknown parameter θ, is estimated up to an arbitrary degree of numerical accuracy by

$$\bar{h}(\theta) = \sum_{t=1}^{M} h(\theta^{(t)})/M.$$

The marginal posterior distribution can be examined by viewing histogram plots of the MCMC sample or fitting a smoothed kernel density estimate to the sample frequencies.

A widely used approach that reduces the variance of these estimators, especially useful for quantiles and tail area calculations, is known as Rao-Blackwellization, as it is derived from application of the Rao-Blackwell Theorem. It can be shown that if the posterior conditional on some other parameter in the model, ϕ, can be evaluated using the MCMC samples for both θ and ϕ, the estimator

$$\bar{h}_{\phi}(\theta) = \sum_{t=1}^{M} \sum_{j=1}^{M} h(\theta^{(t)}) p(\theta^{(t)} | \phi^{(j)}, y)/M$$

dominates the unconditional sample estimator defined previously, $\bar{h(\theta)}$, in terms of variance (and squared error loss).

6 MCMC Inference for the SDM with marginal augmentation

For the Student-t SDM, as given by (1), the Gibbs sampler can be slow to converge because of posterior dependence among the variance parameters σ_ϵ^2 and Λ. Paradoxically, adding an additional parameter can improve the speed of convergence of the Markov chain simulation. This marginal augmentation or parameter expansion is a technique developed by Meng and Van Dyk (1999) to improve the rate of convergence of the MCMC algorithm. The idea is to reduce the correlation between draws via a working parameter that is not part of the original observed data model. Unlike conditional augmentation, where the working parameter is fixed at a specific value, marginal augmentation minimizes the augmented information by marginalizing over the working parameter. Note that not introducing a working parameter is, in fact, implicitly conditioning on a specific value. Avoiding this conditioning by modeling and integrating out that working parameter can increase the variability in the augmented data and thus reduces the augmented information. Data augmentation and parameter expansion methods dramatically increase the generality and applicability of this approach.

Focusing on the Student-t SDM defined in (1), convergence can be very slow between the homoskedastic variance σ_ϵ^2 and the heteroskedastic term Λ. If a posterior draw for σ_ϵ^2 is close to zero, then the draw for Λ will also be sampled with values near zero, and so on. Following Meng and Van Dyk's parameter expansion approach, we can reduce the correlation by adding a new working parameter whose only role is to allow the Gibbs sampler to move

in more directions and thus improve the convergence. To accomplish this, we rewrite (1) as

$$\begin{aligned} y &= \rho W y + \iota_n \alpha + X\beta + WX\gamma + \frac{\sigma_\epsilon^2 z}{\sqrt{\Lambda}} \\ z &\sim N(0, I_N) \\ \Lambda &\equiv diag(\lambda_1, \ldots, \lambda_n) \\ \lambda_i &\sim \chi^2_\nu/\nu. \end{aligned} \tag{7}$$

The expanded model is

$$\begin{aligned} y &= \rho W y + \iota_n \alpha + X\beta + WX\gamma + \frac{\sqrt{\omega}\sigma_\epsilon^2 z}{\sqrt{q}} \\ z &\sim N(0, I_N) \\ q &\equiv diag(q_1, \ldots, q_n) \\ q_i &\sim \omega\chi^2_\nu/\nu. \end{aligned} \tag{8}$$

The parameter $\omega > 0$ can be viewed as an additional scale parameter. In this new specification, q plays the role of $\omega\Lambda$. Thus, introducing ω does not alter the model we are fitting.

Note that since $q_i = \omega\lambda_i$, then λ_i corresponds to q_i when $\omega = 1$. We expect marginal augmentation with a working prior independent of $\theta = (\beta, \rho, \sigma_\epsilon^2)$ to improve the rate of convergence. We choose $p(\omega)$ to be the proper conjugate prior for ω in $p(Y, \lambda|\theta, \omega)$, namely, $\delta\chi_\gamma^{-2}$, where $\delta > 0$, $\gamma > 0$. As in Meng and van Dyk (1999), we use the standard improper prior $p(\beta, \log\sigma_\epsilon^2) \propto 1$. Under this prior, Geweke (1993) shows that the posterior mean and standard deviation exist only if the prior $p(\nu)$ is null in the interval 0 to 4. We will assume the latter prior to be exponential $p(\nu) = \exp(\nu_0)$. The MCMC algorithm for this expanded model has the following steps:

(a) $q_i|\beta, \sigma_\epsilon^2, \rho, \omega, Y \sim \frac{\omega}{(y_i - \rho\sum_j^n w_{ij}y_j - x_i\beta)^2/\sigma_\epsilon^2 + \nu}\chi^2_{\nu+1}$ independently for $i = 1, \ldots, n$;

(b) $\omega|Y, q \sim \frac{\delta + \nu\sum_{i=1}^n q_i}{\chi^2_{\gamma+n\nu}}$;

(c) $\beta|\sigma_\epsilon^2, Y, q, \rho, \omega \sim N(c, T)$, where $c = (X'qX)^{-1}X'q(I_n - \rho W)y$ and $T = \omega\sigma_\epsilon^2(X'qX)^{-1}$;

(d) $\sigma_\epsilon^2|Y, q, \omega \sim \frac{\sum_{i=1}^n q_i(y_i - \rho\sum_j^n w_{ij}y_j - x_i\beta)^2}{\omega\chi^2_{n+1}}$;

(e) $\rho|\beta, \sigma_\epsilon, q \propto |I_n - \rho W| \exp\{-\frac{1}{2\sigma_\epsilon^2}e'qe\}$, where $(I_n - \rho W)y - X\beta$

(f) $\nu|y, q \propto \left(\frac{\nu}{2}\right)^{\frac{N\nu}{2}} \Gamma\left(\frac{\nu}{2}\right)^{-N} \exp(\eta\nu)$, with $\eta = \frac{1}{\nu_0} + \frac{1}{2}\sum_{i=1}^N [\ln(q_i^{-1} + q_i)]$

The first four steps involve Gibbs sampling, but the last two posterior densities are non-standard, and a Metropolis-Hastings step is implemented. A random walk MH algorithm with a normal increment random variable is therefore used for these steps, as described in section 4.2.

Given these conditional distributions, we can implement a data augmentation algorithm with marginal augmentation. At iteration $(t+1)$, we draw $q^{(t+1)}$ from the conditional with marginal augmentation,

$$p(q|\beta, \sigma_\epsilon^2, \rho, Y) = \int p(q|\mu, \sigma_\epsilon^2, Y, \omega)p(\omega)d\omega. \tag{9}$$

The implementation of this marginal augmentation is performed using the following scheme:

(a) Step 1 Draw ω from its prior $p(\omega)$ and then q from $p(q|\beta, \sigma_\epsilon^2, Y, \omega)$.

(b) Step 2 Given q, $(\beta, \sigma_\epsilon, \rho)$ is generated from $p(\beta, \sigma_\epsilon, \rho|Y, q) = \int p(\beta, \sigma_\epsilon, \rho|Y, q)p(\omega|Y, q)d\omega$ by first drawing ω from the posterior $p(\omega|Y, q)$, then drawing β, σ_ϵ and ρ given ω their posterior distributions.

As a comparison, the conditional augmentation approach would fix $\omega = 1$ and ignore its posterior distribution.

6.1 Monte Carlo Results

A Monte Carlo experiment was conducted to evaluate the performance of the above sampling method and compare the conditional vs. the marginal augmented methods. The data generating process is shown in (10).

$$\begin{aligned} y &= \rho W y + \iota_n \alpha + X\beta + WX\gamma + \frac{\sigma_\epsilon^2 z}{\sqrt{\Lambda}} \\ z &\sim N(0, I_N) \\ \Lambda &\equiv diag(\lambda_1, \ldots, \lambda_n) \\ \lambda_i &\sim \chi_\nu^2/\nu \end{aligned} \tag{10}$$

The spatial weight matrix W was generated using random points in conjunction with the Matlab Delaunay routine to produce a symmetric contiguity weight matrix that is then row-normalized (see Chapter 4 in LeSage and Pace, 2009). Explanatory variables x_{it} were generated from zero mean independent normal distributions with a variance of four $(N(0, 4))$. A discussion about the impact of the choice of the hyperparameters δ and γ for the prior distribution of ω can be found in Meng and van Dyk (1999). We set $\gamma = 2$ and $\delta = 0.001$. Table 1 presents posterior means, standard deviations and numerical standard error (NSE) measures of the accuracy associated with estimates based on marginal versus conditional data augmentation steps.

The marginal augmentation is less sensitive to the choice of starting value and decreases the numerical standard errors. Conditioning on the working parameter $\omega = 1$, reduces the speed of convergence of the chain. The parameter σ_ϵ^2 needs clearly more iterations to be completely independent from the initial iterations under the conditional marginal data augmentation.

7 Spatio-temporal Model

In this section, the SDM model with heteroskedasticity is extended into a dynamic panel data model that accommodates spatial dependence. A variety of models that control for serial correlation and spatial dependence across locations have been explored (see Lee and Yu 2010; for a complete review). Yu et al. (2008) analyse a specification that allows for both time and spatial dependence as well as a cross-product term reflecting spatial dependence at a one-period time lag. This last term can be interpreted as spatial diffusion that takes place over time. Parent and LeSage (2009) extend this approach, introducing a space-time filter that can be applied to the dependent variable or the error term. This filter implies a constraint on the mixing term that reflects spatial diffusion or space-time covariance. Parent and LeSage (2009) show that this constraint allows for a number of simplifications in the Bayesian MCMC estimation scheme.

The space-time filter is applied to the following panel data model with random effects and heteroskedastic disturbances across locations:

$$\begin{aligned} y_t &= \iota_N \alpha + x_t \beta + W x_t \gamma + \eta_t \quad t = 0, \ldots, T \\ \eta_t &= \mu + \frac{\sigma_\epsilon^2 z_t}{\sqrt{\Lambda}} \\ z_t &\sim N(0, I_N) \\ \Lambda &\equiv diag(\lambda_1, \ldots, \lambda_n) \\ \lambda_i &\sim \chi_\nu^2 / \nu \end{aligned} \tag{11}$$

where $y_t = (y_{1t}, \ldots, y_{Nt})'$ is the $N \times 1$ vector of observations for the tth time period, α is the intercept, ι_N is an $N \times 1$ column vector of ones, x_t denotes the $N \times k$ matrix of non-stochastic regressors and μ is an $N \times 1$ column vector of random effects, with $\mu_i \sim N(0, \sigma_\mu^2)$. The random error terms $\varepsilon_t = \sigma_\epsilon^2 z_t / \sqrt{\Lambda}$ are assumed to be independent and identically distributed with zero mean and a variance $\sigma_\varepsilon^2 \Lambda^{-1}$. We make the traditional assumption that μ is uncorrelated with ε_t, and Λ represents the heteroskedastic covariance matrix.

To define the time filter, let C be the Prais-Winsten transformation, where ϕ is the autoregressive time dependence parameter. This filter is defined as:

$$C = \begin{pmatrix} \psi & 0 & \ldots & 0 \\ -\phi & 1 & \ldots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \ldots & -\phi & 1 \end{pmatrix}. \tag{12}$$

Specification of ψ, the (1,1) element in C depends on whether the first period is modeled or assumed to be known. For simplicity, we will assume the first observations to be known. The space filter is defined as a nonsingular matrix $B = (I_N - \rho W)$, where ρ is a scalar spatial dependence parameter and W is the known $N \times N$ spatial weight matrix as above. The proposed space-time filter then corresponds to the Kronecker product of the matrices C and B,

$$C \otimes B = I_{N,T+1} - \rho I_{T+1} \otimes W - \phi L \otimes I_N + (\rho \times \phi) L \otimes W, \tag{13}$$

where L is a $(T+1) \times (T+1)$ matrix based on the time-lag operator. This filter implies a restriction that the parameter associated with spatial effects from the previous period $(L \otimes W)$ is equal to $-\rho \times \phi$. Parent and LeSage (2009) show that applying this space-time filter to the dependent variable greatly simplifies the estimation procedure when an optimal predictor is used to model the first observation. They also advocate that imposing this constraint would simplify the interpretation of the marginal effects.

We decide to ignore the issues pertaining to prediction of the first period cross-section values, and apply the filter to the dependent variable resulting in a model specification:

$$\begin{aligned} (C \otimes B)y &= \iota_{N,T+1}\alpha + x\beta + (I_{T+1} \otimes W)x\gamma + \eta \\ \eta &\sim N(0, \tilde{\Omega}) \end{aligned} \tag{14}$$

where $y = (y_0', \ldots, y_T')'$, $x = (x_0', \ldots, x_T')'$ and

$$\tilde{\Omega} = \sigma_\mu^2 (J_{T+1} \otimes I_N) + \sigma_\varepsilon^2 I_{T+1} \otimes \Lambda_N^{-1}, \tag{15}$$

with $J_{T+1} = \iota_{T+1}\iota_{T+1}'$.

A number of studies have treated the parameter $\rho \times \phi$ associated with the cross-product term in different ways. Anselin (1988) proposed a related "time-space dynamic model" specification explored by Yu et al. (2008) who relaxed the implied constraint $\theta = -\rho \times \phi$ and estimated an unrestricted parameter θ. We will start with this general specification and show that the constraint implied by the space-time filter is relevant and makes the model easy to interpret. Since we ignore the first period, the general panel data model specification with random effects for $t = 1, \ldots, T$, is given by,

$$\begin{aligned} y_t &= \phi y_{t-1} + \rho W y_t + \theta W y_{t-1} + \iota_N \alpha + x_t \beta + W x_t \gamma + \eta_t \\ \eta_t &= \mu + \varepsilon_t. \end{aligned} \tag{16}$$

One advantage of the Bayesian MCMC scheme we propose for this specification is that it does not require integration over the random effects that appear in the likelihood. However, integration over these parameters can reduce serial dependence in the MCMC samples of parameters drawn. A formal expression of posterior distributions for this specification can be found in Parent and LeSage (2009). The only difference relies on the heteroskedastic term λ_i that is generated from the following Chi-Square distribution,

$$\lambda_i | \beta, \sigma_\epsilon^2, \rho, \omega, Y \sim \frac{\chi_{\nu+T}^2}{\sigma_\epsilon^{-2} e_i' e_i + \nu} \quad i = 1, \ldots, n, \tag{17}$$

where $e_i = y_i - \phi y_{i,-1} - \rho \sum_{j=1}^N w_{i,j} y_j - \theta \sum_{j=1}^N w_{i,j} y_{j,-1} - \alpha \iota_T - x_i \beta - \sum_{j=1}^N w_{i,j} x_j \gamma$, $y_i = (y_{i,1}, \ldots, y_{i,T})'$ and $y_{i,-1} = (y_{i,0}, \ldots, y_{i,T-1})'$. We also set the hyperparameter $\nu = 4$, consistent with our prior belief in heteroskedasticity.

7.1 Empirical Application

In this empirical illustration, we model and estimate the presence of interregional technological dependence. We rely on a simple model of semi-endogenous growth developed by

Jones (2002). This empirical analysis shows that implementing spatial and time dependence conveys important information regarding to what extent innovative activities spill over to neighboring states.

Based on the model described by Jones (2002), we propose a dynamic specification where the stock of knowledge in the neighboring regions has spillover effects on the growth rate of ideas in region i,

$$\frac{\dot{A}_i(t)}{A_i(t)} = \delta L_i(t)^{\lambda} A_i(t)^{\gamma-1} \prod_{j \neq i} A_j(t)^{\psi w_{ij}}. \tag{18}$$

According to equation (18), the number of new ideas produced at any point in time is driven by the number of researchers and the existing stock of ideas in region i as well as in neighboring regions. The parameter λ represents the effect of research on new ideas and allows for the possibility of duplication. For now, we assume $|\gamma| < 1$ and $|\psi| < 1$, but stability conditions are discussed in more detail later. Based on Parent (2011), we define the connectivity structure W using a measure of both geographical as well as technological proximity. The spatial weight scheme is based on the concept of $5-$nearest neighbors where these 5 neighbors will receive varying weights based on the measure of technological proximity.

Parent (2011) shows that using the log-linearization of the equation (18) around the vector of steady state growth rates g_A where $A(t)$ and $L(t)$ are growing at constant rates, corresponds to:

$$\begin{aligned} \frac{\dot{A}_i(t)}{A_i(t)} &= g_A(1 - \log(g_A/\delta)) + \\ & \quad g_A \left[\lambda \log(L_i(t)) - (1-\gamma)\log(A_i(t)) + \sum_{(j \neq i)} \psi w_{ij} \log(A_j(t)) \right] \end{aligned} \tag{19}$$

And we can rewrite (19) as

$$\log(A_i(t+1)) = \phi \log(A_i(t)) + \sum_{(j \neq i)} \theta w_{ij} \log(A_j(t)) + \alpha + \beta \log(L_i(t)), \tag{20}$$

where $\phi = -g_A(1-\gamma) + 1$, $\theta = g_A\psi$, $\alpha = g_A(1 - \log(g_A/\delta))$ and $\beta = g_A\lambda$.

The parameter θ captures the impact of accessible external ideas on regional innovative activities also called interregional knowledge spillovers. We can add to the econometric specification problems noted by Jones (2002) omitted variables bias that would arise from excluding $\sum_{(j \neq i)} \theta w_{ij} \log(A_j(t))$ from the model by assuming $\psi = 0$, leading to $\theta = g_A\psi = 0$.

We extent the theoretical framework (20), where the diffusion process is similar to an autoregressive model where spatial interaction occurs with a lag of one period. We introduce the traditional simultaneous spatial lags used in cross-sectional models from the spatial econometrics literature, where the right-hand-side variable takes the form: $\sum_{(j \neq i)} \rho w_{ij} \log(A_j(t+1))$.

The stock of patents per capita for state i at the period t results in the following regres-

sion:

$$\begin{aligned} \log(A_{it}) &= \alpha + \sum_{(j \neq i)} \rho w_{ij} \log(A_{j,t}) + \phi \log(A_{i,t-1}) + \sum_{(j \neq i)} \theta w_{ij} \log(A_{j,t-1}) \\ &\quad + \beta \log(L_{i,t-1}) + \gamma W \log(L_{i,t-1}) + \eta_{it} \\ \eta_{it} &= \mu_i + \varepsilon_{it}. \end{aligned} \tag{21}$$

Externalities generated by one region are allowed to influence neighboring regions within the same (annual) time period (the spatial effect), the same region in subsequent periods (the time effect), as well as neighboring regions in subsequent periods (the space-time diffusion effect). This space-time dynamic allows us to compare the relative importance of contemporaneous spatial dependence with time dependence and spatial interaction from the previous periods.

To estimate this model, we must measure the stock of ideas. Observable measures of new ideas at a regional or international level are never perfect. We organize the analysis by focusing on observed number of U.S. domestic patents, a useful indicator of the state level of realized innovation for a given period. We estimate the knowledge production function using a dataset on patenting activity and its determinants covering the period 1994 to 2005 and 49 states.[3] The data include patents granted per capita for each state in each year along with measures of the factor inputs in the production function for ideas/knowledge.

Skilled labor $L_i(t)$ for each state i at time period t is measured using two explanatory variables, $doc_i(t)$, the number of doctoral recipients, and $expRD_i(t)$, total research and development expenditures as a percentage of gross state product. Total R&D expenditures are calculated by adding all sources of funds: industry, public and private non-profit institutes and universities.

7.2 Estimation results

Estimation results are presented in Table 2, based on a sample of 50,000 draws collected after a burn-in period of 10,000 draws. In the following discussion of the parameter estimates we relied on 5 and 95 percentage points of the highest posterior density intervals (HPDI) to draw inferences regarding whether the posterior means were different from zero.

As explained in Pace and LeSage (2009), low levels of spatial dependence between neighboring regions can over time lead to a significant amount of inter-connectivity between regions in the long-run knowledge production process. Ignoring the low levels of observed spatial dependence will have dramatic impacts on the long-run estimates and inference regarding the regional knowledge production and diffusion process.

Traditionally, a positive effect of spatial dependence is interpreted as local spillover effects related to the presence of knowledge stocks in neighboring regions. Parent and LeSage (2008) make the point (in the context of European regions) that positive spatial dependence of this type may arise when regions possess the ability to absorb and to adopt new technologies of their neighbors. Further, R&D activities can increase the incidence of technology diffusion by enhancing a region's absorptive capacity. Positive spatial dependence found here using

[3]The District of Columbia is treated as a state and the states of Alaska and Hawaii are omitted.

the space-time model leads to an inference that R&D expenditures will directly increase the level of innovation occurring in a region over time.

In fact, as explained by Debarsy, Ertur and LeSage (2010), a change to explanatory variable r at time t will have direct and indirect impacts on the own- and other-region dependent variable values at time t, as well as impacts on both own- and other-regions in future time- periods. This diffusion over space as time passes arises when the model includes non-zero time dependence captured by the parameter ϕ.

Turning to the restriction implied by the space-time filter $\theta = -\rho \times \phi$, estimation results presented in Table 2 reveal that this restriction is consistent with the data for both specifications. The partial derivatives for this situation are shown in (22) for the case where we change the explanatory variable $x_1^{(r)}$ at time period 1, and measure the impacts at a one- through t-period horizon. Since the estimation results confirm the time-separability constraint $\theta = -\rho \times \phi$, the partial derivative can be rewritten as,

$$\frac{\partial y_t}{\partial x_1^{(r)}} = (\phi^{t-1} + \phi^{t-2} + \ldots + \phi + 1)B^{-1}(I_N\beta_r + W\gamma_r). \quad (22)$$

This greatly simplifies interpretability of the dynamic responses for any number of time periods. Given estimates for the parameters β_r, ρ and ϕ , we can easily calculate dynamic responses for any number of time periods. In fact, the diffusion over time and space takes the form of time-discounting based on the time-dependence parameter ϕ of the contemporaneous spatial effects captured by the $N \times N$ matrix B^{-1}.

Table 2 shows scalar summary measures of the effects estimates for spatial dependence ($\rho = 0.42$) that is relatively weaker than time dependence $\phi = 0.92$, which leads to larger time and space-time diffusion effects relative to the spatial effects. Based on the stationary conditions defined by Parent and LeSage (2009), the process is stationary since $\phi + \rho + \theta < 1$ and $\phi - (\rho - \theta) > -1$.

Table 3 reports cumulative spatial effects decomposed into direct, indirect and total effects. The direct effects correspond to own-partial derivatives that measure the impact on region i from changes in the explanatory variable value of region i. However, these include some feedback impacts discussed in LeSage and Pace (2009), since changes in region i influence the neighbors and region i is in turn influenced by its neighbors. The indirect effects are cumulated over neighboring spatial regions, and correspond to the cross-partial derivatives, and the final column shows the total effects which is the sum of the direct and indirect effects. In our model, the spatial effects are separable from the time effects, and these do not change over time since the spatial configuration of the regions remains the same and we restrict the spatial dependence parameter to be fixed over all time periods. The differences between the cumulative total effects and the spatial effects reflect the importance of the time effects. In the case R&D expenditures, we see a 0.5096 direct cumulative effect value and a direct spatial effect of 0.0632, so the difference of 0.4464 represents cumulative direct time effects (which we calculated over a 14 year horizon). In comparison with the coefficient estimate of 0.0602 from Table 2 for this variable, the direct effects estimate reported in Table 3 includes a feedback loop that arises in our space-time dynamic panel model.

Consistent with the ideas-based growth literature, the results suggest that the level of innovation is positively influenced by the level of effort devoted to the ideas sector. Expen-

ditures on R&D have a more permanent impact on the growth process if a highly skilled labor force eases the adoption of new technologies. Of course, this is consistent with the observation that regions with advanced levels of technology often have strong links with education, especially at the doctoral level. Thus, more education should lead to higher rates of technological progress via improvements in labor force quality. However for both models, the effect of the variable *LDoc* is not statistically significant.

As shown in Parent (2011), these results confirm that interactions between regions are spatially limited and localized spillovers effects can lead to regional clusters with persistently different levels of innovative activity.

8 Conclusion

This chapter shows how the Bayesian approach provides a complete inferential tool-kit for a variety of cross sectional and panel data spatial models. Bayesian methods have recently produced some remarkably efficient solutions to complex infrence problems. The approach is based on a combination of hierarchical prior modeling and MCMC simulation methods. Interestingly, this approach is able to tackle estimation and model interpretation in situations that are quite challenging by other means.

Marginal data augmentation improves the convergence properties of the MCMC sampler. This method expands the parameter space with a working parameter that is only identifiable given the augmented data. Placing a prior distribution directly on the identifiable parameters results in enormous computational gain. This prior specification can make the model easier to estimate and interpret in many complex cases like multivariate and multinomial discrete choice models.

While this chapter is necessarily too brief to provide a self-contained guide, hopefully it sheds enough light on the main conceptual issues to demonstrate that using Bayesian MCMC inferential tools allows for broad generality in model specification, and is relatively simple to use in practice. The growth of Bayesian MCMC spatial econometric methods continues at a rapid pace as the Bayesian approach becomes more widely understood, and as software and computing power become more readily available.

References

Anselin L (1988) Spatial econometrics: Methods and models Boston MA, Kluwer Academic

Autant-Bernard C, LeSage JP (2011) Quantifying Knowledge Spillovers Using Spatial Econometric Models. Journal of Regional Science 51:471-496

Banerjee S, Carlin B, Gelfand, A (2004) Hierarchical Modeling and Analysis for Spatial Data. Chapman & Hall, Boca Raton

Barry R, Pace RK (1999) Monte Carlo estimates of the log determinant of large sparse matrices. Linear Algebra and its Applications 289:41-54

Chib S, Greenberg E (1995) Understanding the Metropolis-Hastings Algorithm. The American Statistician 49:327-335

Chib S (2008) Panel Data Modeling and Inference: A Bayesian Primer. The Econometrics of Panel Data. (eds L. Matyas and P. Sevestre). 479-515. Springer-Verlag, Berlin Heidelberg

Debarsy N, Ertur C, LeSage, JP (2012) Interpreting dynamic space-time panel data models. Statistical Methodology 9:158-171

Ertur C, Koch W (2007) The Role of Human Capital and Technological Interdependence in Growth and Convergence Processes: International Evidence. Journal of Applied Econometrics 22:1033-1062

Gamerman D, Lopes HF (2006) Markov Chain Monte Carlo. Chapman & Hall

Gelfand, AE, Smith AFM (1990) Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association 85:98-409

Gelman A, Carlin JB, Stern HS, Rubin DB (2004) Bayesian Data Analysis, Second Edition. Chapman & Hall

Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence 6:721-741

Geweke J (1993) Bayesian treatment of the independent Student-t linear model.Journal of Applied Econometrics 8:519-540

Geyer C (2011) Introduction to Markov Chain Monte Carlo. Handbook of Markov Chain Monte Carlo, Eds. S. P. Brooks, A. Gelman, G. Jones and X.-L. Meng. Chapman and Hall/CRC Press

Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97-109

Jones CI (2002) Sources of U.S. Economic Growth in a World of Ideas. American Economic Review 92:220-239

Kakamu KW, Polasek W, Wago H (2011) Production technology and agglomeration for Japanese prefectures during 1991-2000. Papers in Regional Science forthcoming

Lee LF, Yu J (2010) Some recent developments in spatial panel data models. Regional Science and Urban Economics 40:255-271

Lesage JP, Fischer MM (2008) Spatial Growth Regressions: Model Specification, Estimation and Interpretation, Spatial Economic Analysis 3:275-304

LeSage JP, Pace RK (2009) An introduction to spatial econometrics. CRC Press, Boca Raton [FL]

Meng XL, van Dyk DA (1999). Seeking efficient data augmentation schemes via conditional and marginal augmentation. Biometrika 86:301-320

Metropolis N, Rosenbluth, AW, Rosenbluth, MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machine. Journal of Chemical Physics 21:1087-1092.

Metropolis N, Ulum S (1949) The Monte Carlo Method. Journal of the American Statistical Association 44:335-341

Parent O (2011) A Space-Time Analysis of Knowledge Production. Journal of Geographical Systems (forthcoming)

Parent O, LeSage JP (2008) Using the variance structure of the conditional autoregressive spatial specification to model knowledge spillovers. Journal of Applied Econometrics 23:235-256

Parent O, LeSage JP (2009) Spatial dynamic panel data models with random effects, working paper

Tanner MA, Wong W (1987) The calculation of posterior distributions by data augmentation (with discussion), Journal of the American Statistical Association 82:528-550

Tierney L (1994) Markov Chains for Exploring Posterior Distributions (with discussion), Annals of Statistics, 22:1701-1762

Wang X, Kockelman K, Lemp J (2011) The Dynamic Spatial Multinomial Probit Model: Analysis of Land Use Change Using Parcel-Level Data, working paper

Yu J, de Jong R, Lee LF (2008) Quasi-Maximum Likelihood Estimators For Spatial Dynamic Panel Data With Fixed Effects When Both n and T Are Large. Journal of Econometrics 146:118-134

Tables

Table 1: Monte Carlo simulations for $n = 500$ and 1,000 iterations

		Marginal DA, $\omega \sim \delta\chi_\gamma^{-2}$			Conditional DA, $\omega = 1$		
Parameter	True value	Mean	s.d.	NSE	Mean	s.d.	NSE
ρ	0.7	0.6957	0.0086	0.6785	0.6977	0.0046	0.6882
β	1	1.0011	0.0096	0.9820	1.0020	0.0100	0.9820
γ	1	0.9722	0.0266	0.9820	1.0492	0.0352	0.9838
σ_ε^2	1	0.9979	0.7109	0.3096	0.9434	0.7515	0.2782
ν	4	4.0224	0.0111	4.0015	4.0456	0.0423	4.0015

Table 2: Estimation results

Parameters	Post. mean	s. d.	Lower 0.05	Upper 0.95
Constant	0.2949	0.1403	0.0949	0.5444
doc	0.0028	0.0130	-0.0161	0.0274
expRD	0.0602	0.0256	0.0214	0.1046
Wdoc	-0.0260	0.0168	-0.0587	0.0006
WexpRD	-0.0354	0.0268	-0.0825	0.0040
ρ	0.4157	0.0353	0.3607	0.4807
ϕ	0.9152	0.0632	0.7320	0.9657
θ	-0.3798	0.0413	-0.4477	-0.2999
σ_μ^{-2}	0.0102	0.0113	0.0012	0.0354
σ_ε^{-2}	0.0145	0.0013	0.0125	0.0168

Table 3: Scalar summary estimates of the R&D effects

	Lower 0.05	Median expRD	Upper 0.95
Spatial Effects			
Direct	0.0225	0.0632	0.1099
Indirect	0.0141	0.0398	0.0692
Total effects	0.0366	0.1030	0.1790
Cumulative Effects			
Direct	0.1812	0.5096	0.8855
Indirect	0.1140	0.3208	0.5574
Total effects	0.2952	0.8304	1.4429