Wolf, L. J., Anselin, L., & Arribas-Bel, D. (2018). Stochastic Efficiency of Bayesian Markov Chain Monte Carlo in Spatial Econometric Models: An Empirical Comparison of Exact Sampling Methods. *Geographical Analysis*, *50*(1), 97-119. https://doi.org/10.1111/gean.12135

Stochastic Efficiency of Bayesian MCMC in Spatial Econometric Models:

An Empirical Comparison of Exact Sampling Methods[*]

Levi John Wolf — ljw2@asu.edu

Luc Anselin — anselin@uchicago.edu

Daniel Arribas-Bel — D.Arribas-Bel@liverpool.ac.uk

May 15, 2017

Abstract

Spatial econometric specifications pose unique computational challenges to Bayesian analysis, making it difficult to estimate models efficiently. In the literature, the main focus has been on extending Bayesian analysis to increasingly complex spatial models. The stochastic efficiency of commonly used Markov Chain Monte Carlo samplers has received less attention by comparison. Specifically, Bayesian methods to analyze effective sample size and samplers that provide large effective size have not been thoroughly considered in the literature. Thus, we compare three MCMC techniques: the familiar Metropolis-within-Gibbs sampling, Slice-within-Gibbs sampling, and Hamiltonian Monte Carlo. The latter two methods, while common in other domains, are not as widely encountered in Bayesian spatial econometrics. We assess these methods across four different scenarios in which we estimate the spatial autoregressive parameter in a mixed regressive, spatial autoregressive specification (or, spatial lag model). We find that off-the-shelf implementations of the newer high-yield simulation techniques require significant adaptation to be viable. We further find that the effective size is often significantly smaller than the nominal size for samples of the spatial autoregressive parameter. In addition, we find that stopping simulation early may result in understated posterior credible interval widths if effective sample size is small. More broadly, we suggest that simulation stopping rules and sample information deserve more attention in both applied and basic Bayesian spatial econometric research.

[*] This research was funded in part by NIH Award 7R01CA1266858-06, GeoSpatial Factors and Impacts II. An earlier version was presented at the 63rd Annual North American Meetings of the Regional Science Association International, November 9-12, 2016. Comments from the participants are greatly appreciated, as are suggestions on earlier drafts by Sergio Rey.

1 Introduction

Bayesian spatial econometrics is undergoing a complex period of transition. Fueled by the rapid growth and subsequent deepening of Markov Chain Monte Carlo (MCMC) methods and theory (Robert and Casella, 2011), an interest in large multilevel models in quantitative social research (Gelman and Hill, 2006), and a longstanding focus on modeling spatial dependence and heterogeneity in econometric systems (Anselin, 1988; Fotheringham and Brunsdon, 1999; Anselin, 2010), Bayesian methods have become increasingly common in spatial econometrics. Early formal treatments of Bayesian estimators for linear models under spatial dependence include Hepple (1979) and Anselin (1980). Subsequent work in the domain has primarily focused on estimating generalized linear models (Smith and LeSage, 2004), or more complex hierarchical specifications (e.g. Dong and Harris, 2015; Lacombe and McIntyre, 2016). Such efforts have been enabled by the veritable explosion in computational power, making it possible to estimate even large hierarchical models using Markov Chain Monte Carlo (Brooks et al., 2011). The Gibbs sampler (Geman and Geman, 1984), leveraged early in spatial statistics (Besag and Green, 1993; LeSage, 1997; Robert and Casella, 2011), is now a widely employed estimation strategy due to its speed, ease of implementation, and flexibility.

Even though the model specifications used in spatial econometrics typically differ from those supported by off-the-shelf automatic Gibbs sampling packages (Spiegelhalter et al., 2007; Finley et al., 2007; Lee, 2013), the proliferation of efficient Bayesian computation techniques has significantly improved the prospects for Bayesian spatial econometric work. However, the notion of "efficiency" is often not straightforward to define in this context. Gibbs sampling made Bayesian estimation of common spatial econometric models simpler to implement, but it did not make the required computations easier. Efforts to improve the complex numerical computations inherent in many spatial econometric specifications are longstanding (Pace and Barry, 1997), primarily focused on the calculation of the determinant and inverse of a large sparse matrix (the Jacobian term in the likelihood function). As it turns out, the concerns of efficient Bayesian computation with the Jacobian term are, in part, similar to those pertaining to Maximum Likelihood estimation for large data sets. Thus, Bayesian estimation has benefited from new techniques developed in Maximum Likelihood contexts, such as fast sparse matrix determinant computations (Smirnov and Anselin, 2001; Pace and LeSage, 2009; Bivand et al., 2013; Bivand and Piras, 2015).

However, there are also unique issues in Bayesian computation, such as prior sensitivity (Gelman, 2006; Polson and Scott, 2012), simulation convergence (Gelman and Rubin, 1992b; Cowles and Carlin, 1996), or posterior propriety (Hobert and Casella, 1996). These are informed by general Bayesian literature and are common to any model estimated in a Bayesian framework. Among these, sampler efficiency is a critical issue for spatial econometric models and is often overlooked. Efficiency, in a broader sense, must be understood to be more than reporting the required burn-in and samples per second of a given sampling technique. Instead, the efficiency of the simulation estimators used in MCMC must include indications of the effective output of a simulation (Flegal et al., 2008). Even for converged chains, simulations typically generate serially correlated draws from the posterior distribution. This means that the effective number of independent samples from the posterior distribution is typically much smaller than the nominal sample size. As in many time series analysis applications, estimates of the variance of a correlated sequence may be understated if this autocorrelation is not accounted for by the estimator.

Thus, we will characterize Bayesian simulation methods using three distinct measures of sampler efficiency: *nominal throughput*, the number of raw samples generated while sampling; *effective throughput*, the number of effectively independent samples generated while sampling; and *yield*, the number of effectively independent samples generated per sample drawn. While significant work has made some Bayesian spatial econometric techniques *numerically* efficient with high nominal throughput, these techniques may not be *stochastically* efficient, and their effective throughput may be unexpectedly low.

In practice, the effective throughput of a simulation is more valuable than the nominal throughout, since the nominal throughput provides no information about quality of the sampler's output. Treating a random series with serial correlation as if it were independent is a well-known issue in time series analysis, and this also applies to the output of MCMC simulation runs. If loss of information due to serial correlation is not addressed or reported, interpretations of posterior samples may be overly precise (Flegal et al., 2008). While simulations with high stochastic efficiency will have good correspondence between the nominal and effective sample sizes, positive serial correlation may result in effective sizes that are significantly lower than the nominal size, exacerbating the potential for overly-precise inference. Recent work in the Bayesian literature has established that effective sample size stopping rules, where simulation is terminated when parameter samples reach a target effective size, are equivalent to older fixed interval width stopping criteria, which stop simulation when posterior credible intervals for parameter estimates reach some specified threshold width (Jones et al., 2006; Flegal, 2008; Gong and Flegal, 2016).

Using effective sample size stopping rules in tandem with thinning, where only a subset of the posterior sample is retained for analysis, posterior inference can accurately match the level of precision contained in the effective sample. In contrast, most Bayesian spatial econometric work uses fixed nominal size stopping rules, where a simulation is run for a fixed number of iterations. Then, some initial fraction of the chain is discarded and the remainder of the chain assessed for convergence failure. If no convergence failure is detected, all draws from the chain after burn-in are used to construct parameter estimates.

In what follows, we examine the commonly-applied Metropolis-within-Gibbs method (LeSage, 1997) for obtaining posterior samples of parameters with two newer alternative sampling methods, Slice-within-Gibbs (Neal, 2003) and a Hamiltonian Monte Carlo algorithm, the No U-Turn Sampler, or NUTS (Hoffman and Gelman, 2014). We restrict our attention to exact sampling algorithms, those that sample directly from the posterior distribution. In doing so, we avoid discussing approximation techniques like Integrated Nested Laplace Approximation (Rue et al., 2009) or gridded Gibbs inversion sampling (Ritter and Tanner, 1992; LeSage and Pace, 2009), that sample from an approximation of the posterior distribution. This focus on exact sampling techniques does not imply that these approximants are somehow sub-par, as these approximation techniques tend to perform quite well relative to exact sampling methods for spatial models (Bivand et al., 2014; Ohtsuka et al., 2015).

With information on yield, we illustrate how the common fixed nominal size stopping rule may result in samples with smaller-than-anticipated effective size. This is contrasted with the performance of an effective size stopping rule. In addition, we demonstrate how simulating a chain for a fixed nominal sample size may provide posterior intervals that understate the true, long-run uncertainty about the parameter estimates, feeding into longstanding concerns about the appropriate length of simulation in Bayesian inference (Gelman and Rubin, 1992b; Raftery and Lewis, 1992b,a; Kass et al., 1998). These demonstrations will consider four scenarios of a mixed regressive, spatial autoregressive (spatial lag) specification, which indicates that more complex model specifications that involve spatial lag terms may inherit this issue. Finally, we conclude by discussing potential methods for improving the numerical and stochastic efficiency of the sampling methods and make recommendations for efficient defaults for sampling and simulation stopping rules based on the results of these experiments.

2 Bayesian Estimation of Spatial Econometric Models

Markov chain Monte Carlo simulation techniques sample from probability distributions by constructing a Markov chain whose stationary distribution is the target probability distribution (Tierney, 1994; Brooks et al., 2011). If this chain can be constructed, then after a sufficient number of steps, the chain will have reached its stationary distribution.[1] This distribution, especially in a Bayesian context, is typically the product of a model likelihood and prior information on the parameters. MCMC techniques aim to sample from this stationary distribution, but the actual samples generated in any single simulation are not guaranteed to converge to the target in a finite number of iterations. Various diagnostics exist to detect convergence failure, but no convergence diagnostic is sufficient to indicate convergence to the target stationary distribution (Cowles and Carlin, 1996).

In addition to uncertainty about convergence, many MCMC techniques are not efficient in generating samples from the target stationary distribution. When the sampling procedure moves slowly through the parameter space, each draw from the target distribution may be highly correlated with the previous draw. Serial correlation leads to a loss of information that can occur regardless of what MCMC technique is used to draw from the posterior, and reflects general issue for stochastic simulation techniques.

Thus, the critical measure of sample size to assess the quality of ergodic estimates in MCMC is the effective sample size, not the nominal size. The effective sample size accounts for serial correlation between draws using an unbiased estimator of the variance in a serially-correlated signal. The effective size of a given sample may be higher or lower than the nominal size, but in many cases, positive serial autocorrelation drives the effective size down significantly. One statement of the effective sample size for a single correlated sequence of length m drawn from a converged Markov process is given by Gelman et al. (2014, §11.5):

$$\hat{m}_e = \frac{m}{1 + 2\sum_i^k \hat{\phi}_i} \tag{1}$$

where $\hat{\phi}_i$ is an estimate of the serial correlation at the ith lag, out to k maximal lags. From a practical perspective, k must be estimated from the chain using standard signal analysis techniques (Lütkepohl and Krätzig, 2004), with Gelman et al. (2014, §11.5) suggesting k as the first lag at which two sequential autocorrelation coefficients are negative. A similar (but strongly consistent) statistic for effective sample size is has been used for a stopping rule, terminating simulation when the effective sample size moves

[1] This is also referred to as the “ergodic,” “steady-state,” or “invariant” distribution,

above a threshold set in advance (Gong and Flegal, 2016). However, in Bayesian spatial econometrics, explicit simulation stopping rules are not commonly used.

In spatial econometrics, Bayesian work has focused extensively on nominal throughput, with samplers designed to attain a pre-specified nominal sample size as quickly as possible. Under this mode of analysis, a chain is simulated for a fixed number of iterations, an initial fraction of draws is discarded as “burn in,” and the remaining draws are assessed for lack of convergence. If no convergence failure is detected, all of the remaining chain is used in the analysis. This “fixed nominal size” stopping rule results in chains whose total number of iterations is predetermined, but whose information content is unknown in practice because no measure of sample information is used to determine when to stop the simulation or remove observations from the trace.

The difference between nominal and effective sample size has been appreciated for some time. For example, in arguably the first application of Gibbs sampling in a spatial econometric context, LeSage (1997) suggested the use of the Raftery and Lewis (1992a) diagnostic to identify a nominal stopping size that accounts for potential serial correlation in simulated draws. This contrasts with the typical work flow discussed in Section 1. Cases where the typical work flow is used include the spatial panel model defined in Parent and LeSage (2010), the simultaneous autoregressive spatial interaction model of LeSage and Llano (2013), the models fit by Jensen et al. (2013) to analyze British elections, or the hierarchical SAR model developed by Dong and Harris (2015). This is also the case for the development of spatial probit models in Smith and LeSage (2004) and extensions or applications in spatially-dependent discrete choice modeling, such as the analysis of land use/land cover change for high-yield rice variants (Holloway et al., 2002), extensions to panel structures (Wang and Kockelman, 2009; Baltagi et al., 2016) or multinomial choice (Wang et al., 2012). As an exception, Dong and Wu (2016) attempt to control for serial autocorrelation through shrinking the sample size by thinning the trace by a factor of ten.

The thorough discussion of Bayesian spatial econometric estimation in LeSage and Pace (2009) acknowledges concerns about “effective draws” in Metropolis-within-Gibbs strategies. But, “effective” is used to refer to the inherent inefficiency of the Metropolis step (Gelman et al., 1996), which only generates new draws from the posterior distribution when a proposal value is “accepted.” The accept/reject issue is a less general problem than loss of information due to correlation between draws (Flegal et al., 2008). The routine use of thinning as a correction for serial correlation is a common folk practice, and issues of effective size and sampler autocorrelation have received attention in the literature (Link and Eaton, 2012; Ohtsuka et al., 2015). Regardless, the most common procedure in Bayesian spatial econometrics is to use a fixed nominal sample size as the stopping criterion and to ignore the effective sample issue.

Much of the Bayesian spatial econometric work in the literature also follows the Metropolis-within-Gibbs approach originally suggested by LeSage (1997).[2] While Gibbs sampling has enabled entirely new model specifications to be fit, contemporary work aims to provide more efficient alternatives (Neal, 1998; Duane et al., 1987; Diaconis et al., 2000; Neal, 2011; Hoffman and Gelman, 2014). This becomes important because the loss in effective sample size may cause the standard errors of parameter estimates to get larger after correcting for autocorrelation in the trace(Flegal, 2008). The worse the simulation yield, the more inflated the nominal sample size will be, so using a correction for positive serial correlation may result in larger standard errors and more uncertainty. This is compounded in specifications that include multiple spatial autoregressive effects, since models with many substantive parameters and deep hierarchical structures are often more difficult to sample efficiently.

A number of high-yield sampling techniques have recently been developed in the Bayesian literature. We consider two in particular: Slice sampling and Hamiltionian/Hybrid Monte Carlo (HMC). Since these may be less familiar, we briefly discuss each in turn.

Slice sampling (Neal, 2003) is a sampling technique designed to reduce random walk behavior in a simulations. It is motivated by the realization that draws of a parameter η from any convex function $f(\eta)$ proportional to the target probability density $P(\eta)$, can be constructed by drawing random pairs (η, h) from underneath the graph of $P(\eta)$. Starting from an initial value η_0, a height h is drawn uniformly from $[0, f(\eta_0)]$. Then, a level set, or *slice*, is constructed as the set of all η_c such that $f(\eta_c) \geq h$. In practice, this slice is constructed by iteratively extending a candidate slice until its endpoints fall outside $P(\eta)$. A draw of η made uniformly at random from this level set is a draw from the probability distribution (Neal, 2003). To illustrate, two iterations of the sampler are shown in Figure 1.

This technique is most efficient when the slice from the previous iteration is likely to span the subsequent iteration as well. Since the first draw below the target distribution at height h is accepted, the probability of an iteration returning the same value as the previous iteration, something common

[2] See LeSage (1997) and LeSage and Pace (2009) for an in-depth discussion of MCMC and Metropolis-within-Gibbs techniques in a spatial econometric context.

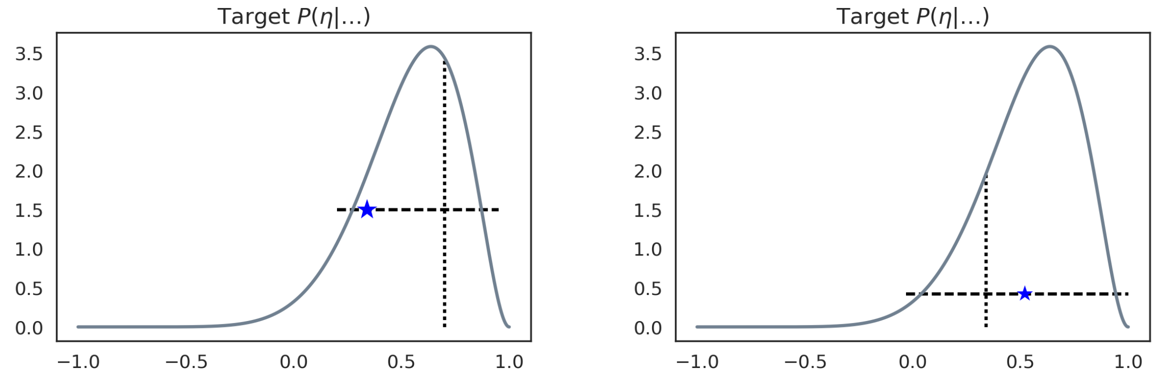

Figure 1: Slice sampling in two frames. On left, the vertical dotted line is the previous draw's value, $\eta_0 = .7$. The horizontal dashed line is the level set at a randomly-chosen height from, $h = 1.5$, and the star, $\eta_1 = .34$, is a draw made uniformly from the slice. In the next iteration, a level set is constructed at a new random $h = .45$ from under $f(\eta_1)$, and a new η_3 sampled from this set.

in Metropolis sampling, approaches zero. Even though the procedure starts centered at the previous draw, the final spanning slice is sampled uniformly, ignoring the information about the previous draw. In practice, however, draws may still be correlated depending on the many potential interactions between parameters during simulation.

A second sampling strategy that works to suppress random walks is Hamiltonian/Hybrid Monte Carlo (HMC). HMC is a simulation technique that has received significant attention for estimating particularly difficult models, such as those in deeply-nested hierarchical models. A thorough development of HMC from first principles is provided by Brooks et al. (2011, Chapter 5), and an effective conceptual introduction is provided by Betancourt (2017). At a glance, HMC uses both the likelihood and gradient information to iteratively construct proposals that have a high probability of being accepted while still achieving ergodicity. A full presentation of the set of approaches common in HMC is beyond the scope of this paper, but a short explanation follows.

HMC methods introduce auxiliary variables to record information about both the "position" of the sampler, the parameters of interest, and the "momentum" of the sampler, information about where to move in order to achieve better proposals. The joint state of the position and momentum of the simulation is recorded by a Hamiltonian, $H(q, p)$, where q describes the position variables and p the momentum variables in a system of moving particles. Typically, H is an additive function describing the total energy of the sampler:

$$H(q, p) = U(q) + K(p) \tag{2}$$

where U is a function modeling the potential energy of the system and K a function modeling the kinetic energy of the system.

For statistical computation, the joint distribution of the quantities of interest (q) and the auxiliary momentum variables (p) is modeled using a Gaussian distribution:

$$P(q,p) = \frac{1}{Z}\exp\left\{-H(q,p)/T\right\} \tag{3}$$

where T and Z are system-specific tuning parameters. T governs the "temperature," or inherent dispersion, and Z is a system-specific normalizing constant to make $P(q,p)$ a proper density function. Since $H(q,p)$ is additive on p and q, the joint density can be expressed as the product of marginals:

$$P(q,p) = \frac{1}{Z}\exp\left\{\frac{-U(q)}{T}\right\}\exp\left\{\frac{-K(p)}{T}\right\} \tag{4}$$

Since q reflects the substantive parameters, $U(q)$ then becomes the negative log-likelihood function of substantive parameters, a numerical summary of the "position" of the sampler. The momentum variables in $K(p)$ are often modeled using a multivariate normal distribution.

Then, in order to sample new (q,p) pairs, a proposal generating scheme is required. HMC approaches iteratively adjust the momentum and position variates in a way that conserves the total energy of the system, $H(q,p)$. The choice of how to iteratively step forward according to sampler momentum and update sampler position may impact simulation quality, and many iterative updating methods exist. One common rule, the "leapfrog rule", updates the position vectors and momentum vectors in each step j by taking $L-1$ discrete steps of size ϵ along the direction ∇U and updating momentum accordingly:

$$\begin{aligned} q(j+1) &= q(j) + \epsilon p(j)/m_i \\ p(j+1) &= p(j) - \nabla U(q)\epsilon \end{aligned} \tag{5}$$

where m_i is a parameter in the kinetic energy function. A half-step for p is made before the first step and after the last step of to yield a full L steps. So, at a high level,t the sampler constructs proposals by iteratively updating the parameter vector and auxiliary variables according to gradient information about the likelihood and auxiliary model for sampler momentum.

In this formulation, ϵ governs the step size of the forward simulation, tuning how large each $\Delta U(q)$

step should be. Once a candidate has been identified at the end of forward simulation, the final candidate state of both position and momentum variables is accepted according to a typical Metropolis-Hastings accept/reject step. However, the acceptance rate is typically much higher for HMC than for Metropolis-Hastings samplers. At a high level, HMC trades simplicity in generating proposals for higher efficiency with those proposals.

Like many MCMC methods, tuning is required to identify efficient sampler parameters, here ϵ and L. One strategy, the "Langevin" strategy, reduces these two free parameters to one by fixing $L = 1$, so each proposal is only $\nabla U(q)\epsilon$ separated from the starting state, and ϵ can be tuned more judiciously (Marshall and Roberts, 2012). A popular technique for tuning ϵ and L simultaneously is provided by the No U-Turn Sampler (NUTS) of Hoffman and Gelman (2014), and is implemented in cutting-edge Bayesian computation packages such as `Stan` (Carpenter et al., 2016) and `PyMC3` (Salvatier et al., 2016). Standard, manually-tuned HMC algorithms are commonly available (Brooks et al., 2011, Chapter 5), and have also been used in a wide variety of applications.

HMC may be more efficient than other MCMC techniques for hierarchical models, but it requires the evaluation of the likelihood and its gradient possibly multiple times within each iteration. In addition, while HMC has been shown to be more efficient for high-dimensional models with slow-moving parameters, it is less clear whether it is similarly high-yield for low-dimensional but weakly-identified parameters. Finally, HMC is often more complex analytically, since the partial gradient of the likelihood must be derived for any parameter sampled. On this final point, many deployments of HMC for general-purpose modeling use automatic differentiation (Rall and Corliss, 1996; Bergstra et al., 2010; Carpenter et al., 2016). However, as is common with much general-purpose statistical software, the sparse numerical structures of spatial econometric models are typically ignored, so automatic differentiation may yield expressions that are correct, but computationally inefficient to evaluate for spatial models.

3 Model Specification

We compare the three major sampling techniques just reviewed to estimate the parameters in a simultaneous mixed regressive, spatial autoregressive model (spatial lag model)(Anselin, 1988). One way to interpret this specification is that a complete pattern of response Y for N observations is modeled as a function of p covariates X and associated parameters β, Gaussian independent identically-distributed errors ϵ with mean 0 and deviation σ, and an exogenous, predetermined spatial structure matrix, W. The structure matrix W encodes the spatial relationships between each observation and all other observations. Thus, W is typically an $N \times N$ row-standardized matrix constructed before estimation using an appropriate spatial relation for the spatial support of the data. Together, the spatial lag model is commonly expressed in a reduced form:

$$Y = (I - \rho W)^{-1} (X\beta + \epsilon) \tag{6}$$

where I is an $N \times N$ identity matrix. This model has a long history in spatial analysis (Whittle, 1954; Cliff and Ord, 1981), and is chosen here for two reasons. First, many of the more complex hierarchical models depend on efficient sampling of the ρ parameter. This is especially the case in several recently suggested multilevel frameworks with spatial dependence at different levels, such as the hierarchical SAR model (Dong and Harris, 2015), or hierarchical spatial Durbin specifications (Lacombe and McIntyre, 2016).

Second, the likelihood for the model in Eq. 6 poses a unique computational challenge for MCMC. Each time the log-likelihood is evaluated, the log determinant of a large, potentially asymmetric sparse matrix is required. Given the Gaussian ϵ term, Anselin (1988) provides the log-likelihood for the model as:

$$\mathcal{L}(Y|\beta, \sigma^2, \rho) \propto \log\left(|I - \rho W|\right) - \frac{N}{2}\sigma^2 - \left\{\frac{1}{2\sigma^2}(Y - \rho WY - X\beta)'(Y - \rho WY - X\beta)\right\} \tag{7}$$

Because W is often row-standardized, the spatial filter matrix $I - \rho W$ may be asymmetric. If so, typical strategies to evaluate Jacobian determinants, like sparse Cholesky decomposition, are unavailable.

In the spatial econometric literature, considerable work has focused on making this sparse log determinant numerically efficient for maximum likelihood estimation. Importantly, some MCMC techniques require more evaluations of this log determinant in each iteration than others. For example, a typical Metropolis-within-Gibbs step requires exactly two evaluations of this log determinant, since it is evaluated at both the current and proposed ρ value (LeSage, 1997). At best, Slice-within-Gibbs sampling requires at least four evaluations of the likelihood: one to place an upper bound on the height of the level set, at least two when verifying that the slice spans the conditional posterior, and one to ensure the candidate also falls under the conditional posterior. Hamiltonian Monte Carlo algorithms, however, require evaluation at each of the L steps taken during the leapfrog component of an iteration. In addition, the *gradient* of the Jacobian determinant term with respect to ρ is required for HMC. Customized HMC implementations could amortize this cost by computing a single matrix factorization each leapfrog step and using this factorization in both the gradient and likelihood evaluation. However, the automatic differentiation engines that drive HMC in `Stan` or `PyMC3` do not recognize this shared structure. Regardless, the computation of multiple determinants (and inverses) per iteration is expensive, and this expense is amplified if draws are lost to serial correlation or rejected proposals.

With this in mind, we consider four different empirical scenarios to assess the potential for HMC and Slice sampling to improve efficiency in estimating the posterior density of ρ relative to the standard Metropolis-within-Gibbs. Since all methods are consistent and converge, they all recover effectively indistinguishable point estimates. But, as will be discussed below, the posterior intervals do vary between traces and may be impacted by premature stoppage caused by fixed nominal size stopping rules.

In order to control some of the uncertainty involved in the process, we construct both synthetic specifications and the replication of actual empirical models. Two of the specifications pertain to the familiar spatial layout for house prices in Baltimore, MD, initially used by Dubin (1992) and employed in several other replication exercises in the literature. We simulate the dependent variable for the $N = 211$ observations by means of the reduced form model shown in 6. In addition, W is designed as a threshold distance spatial weights matrix, since the data are point addresses of houses. The resulting data generating process for this small-N test case is:

$$Y = (I - .45W)^{-1}\left(2 + -3 * X_1 + -6 * X_2 + 1X_3 + 7X_4 + 5X_5 + \epsilon\right) \tag{8}$$

where ϵ is an independent, normally-distributed error term with $\sigma^2 = 25$, and W is the threshold distance weights matrix. These values are picked arbitrarily to ensure the process is well-behaved during sampling, parameter estimates are well separated, and the ρ parameter is significantly different from 0 at the end of estimation. A second model replicates the hedonic specification of Dubin (1992).

The two remaining examples pertain to the spatial layout of counties in the U.S. South $(N = 1412)$ using queen weights.[3] For the third case, we create a synthetic data set using a subset of covariates

[3] Both the Baltimore and the Southern U.S. counties data sets are part of the PySAL sample data, available from pysal.org. See also Anselin and Rey (2014) for extensive use of these sample data in non-Bayesian spatial econometric analysis.

from a collection of socio-economic data in the context of explaining county-level homicide rates (Baller et al., 2001). The final, fourth case pertains to a model of colorectal cancer screenings explained by demographic covariates (Mobley et al., 2010b,a), where the spatially lagged dependent variable is included to to control for possible spatial feedbacks in the outcome of a policy intervention.

Together, these four cases provide coverage of common spatial econometric problems. First, the small- and large-sized scenarios cover a range of problem sizes common in applied work. In addition, estimating the model over a controlled, well-conditioned dataset and an uncontrolled actual empirical dataset allows us to identify both optimistic and realistic expectations of sampler efficiency for either problem size. The cases from which the empirical examples are drawn, hedonic house price modeling and program analysis in spatial epidemiology, are domains where spatial econometric models are often used, so the designs considered cover a reasonably range of practical estimating conditions. We focus on these empirical examples (or empirically-inspired demonstrations) to show that these effects matter in common applied circumstances.

In all four cases, we use the commonly employed weakly-informative, conditionally conjugate prior choices for the parameters:

$$P(\beta) \sim \mathcal{N}(0, 100) \tag{9}$$

$$P(\sigma^2) \sim IG(.001, .001) \tag{10}$$

$$P(\rho) \sim Unif(-1, 1) \tag{11}$$

While the choice of inverse gamma priors for σ^2 has been called into question in the broader Bayesian statistical literature (Gelman, 2006; Polson and Scott, 2012), we use it here to replicate the designs of the models found in the literature mentioned above. We take intentionally vague for the prior parameters to reflect typical empirical cases where prior information is relatively weak.

4 Results

In all four scenarios, a model was fit using NUTS for each sample step with the `PyMC3` package (Salvatier et al., 2016), which computes the posterior gradients using automatic differentiation in `Theano` (Bergstra et al., 2010). The Metropolis-within-Gibbs described in LeSage and Pace (2009) and a Slice-within-

Gibbs strategy using the Slice sampler described in Neal (2003) were implemented separately. To make throughput comparisons clear, ten thousand samples were drawn first using NUTS, and the sample time recorded.[4] Then, the Slice and Metropolis-within-Gibbs estimators were run for the same amount of time NUTS took to draw $10,000$ samples. Thus, samplers in each case are run for the same amount of time, but the smaller-N case requires a smaller amount of time than the large-N case. NUTS conducts $10,000$ iterations in all cases, but Slice and Metropolis-within-Gibbs will each conduct a different number of iterations during the allotted time.

In all cases, the Metropolis runs were tuned to reach an efficient scale parameter for the Gaussian proposal density that achieved an acceptance rate between $.2$ and $.25$ after burn in. This acceptance rate is in line with the optimal rate derived in Roberts et al. (1997). Since this comparison focuses on stochastic efficiency, this ensures the Metropolis sampling run is operating efficiently and can provide a fair point of comparison. After tuning, the jump size was held fixed for the substantive iterations. Finally, the `CODA` package (Plummer et al., 2006) in `R` (R Core Team, 2016) was used to assess the effective sample size, throughput, and yield of the sample runs according to the effective sample size in Eq. 1.

We focus exclusively on the properties of the ρ traces, since their computation is is expensive and may experience strong serial correlation. The efficiency of the sampler in the ρ step is usually the primary constraint on the overall efficiency attainable for this specification. By extension, this is expected to also hold for more complex models, such as multilevel specifications that contain the spatial autoregressive component.

4.1 Effective Sample Size and Yield

4.1.1 Baltimore Hedonic Model

First, consider the traces for the synthetic Baltimore house price data shown in Figure 2. In this case, the simulation methods have vastly different efficiency. Sampling $10,000$ iterations using NUTS takes

[4] In this comparison, the time taken by the automatic differentiation engine to *compute* the model gradients is omitted to focus explicitly on the efficiency of the sampling technique. It should be noted that the complexity of the Jacobian (and thus the likelihood gradient) depends critically on the density of the weights matrix. In our examples, we have two extremes. The weights for the Baltimore examples are based on a distance criterion, which results in a fairly dense matrix, with an average number of neighbors of about 45 per location and almost 25 percent non-zero links. In contrast, the queen contiguity weights used in the South example are very sparse, with an average of almost 6 neighbors per county and and only 0.4 percent non-zero links (in other words, 99.6 percent of the elements in this matrix are zero). As a consequence, for the Baltimore examples, the automatic differentiation time constituted a significant fraction of overall computing time. For the South examples, sampling time handily dominates gradient derivation time.

$1,028$ seconds, or around 20 minutes. The Metropolis sampler has a much higher throughput, or raw samples per second, than the other two methods. This is shown clearly in the top segment of Table 1. In this case, the effective size of the resulting Metropolis sample, which adjusts for serial correlation between draws, is also larger than that provided by NUTS over the course of sampling. But, the sample yield, or the percent of iterations that generate *independent* samples, is much worse in the Metropolis sample than the other two methods. These two factors come to balance in the effective throughput, placing Metropolis between Slice and NUTS in terms of the number of independent samples drawn per second.

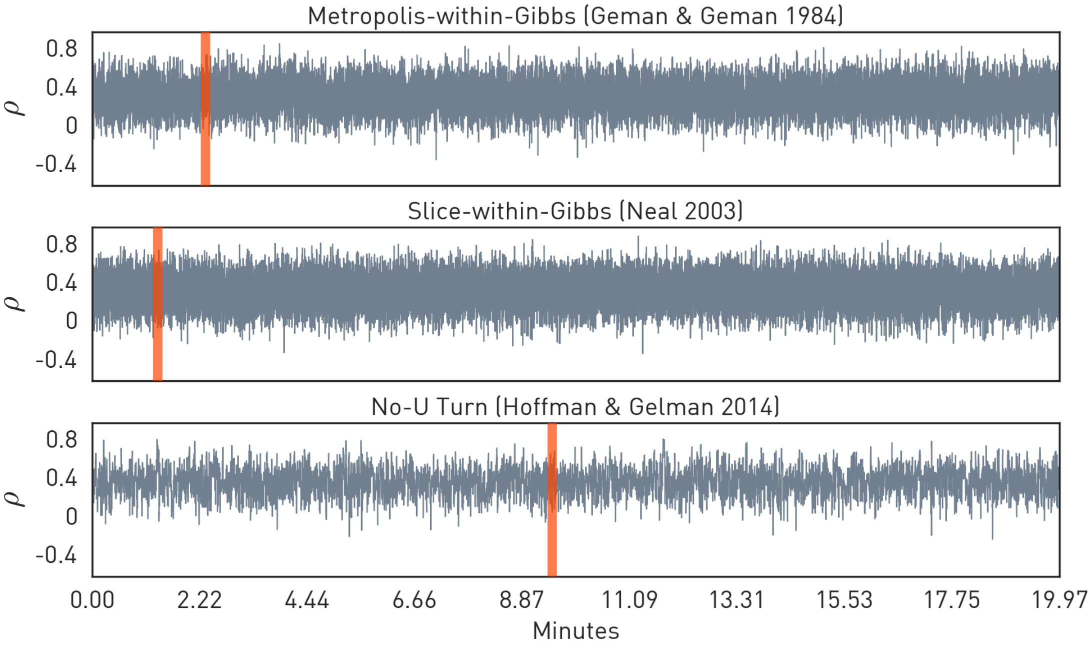

Figure 2: Traces for ρ on the synthetic data model using Baltimore weights, plotted against computation time. The red vertical bar marks the expected time to take 1000 effective samples.

The Metropolis run attains a yield of around 6%, meaning three effectively independent samples are drawn about every fifty iterations. Slice sampling and NUTS both have higher yields than this in the simulated case, and draw an effectively independent sample approximately every fourth iteration in this simulation. However, NUTS takes so long per iteration that the total effective sample size drawn by Metropolis and Slice samplers are larger than the NUTS-generated sample in the allotted time. Using the effective throughput figures, this means that if a fixed effective size stopping rule were used, NUTS would have drawn an effective sample size of $1,000$ in around 8 and a half minutes. In contrast, Slice sampling would have drawn $1,000$ effective samples in nearly a minute and a half. Metropolis would require around 2 and a half minutes to do so.

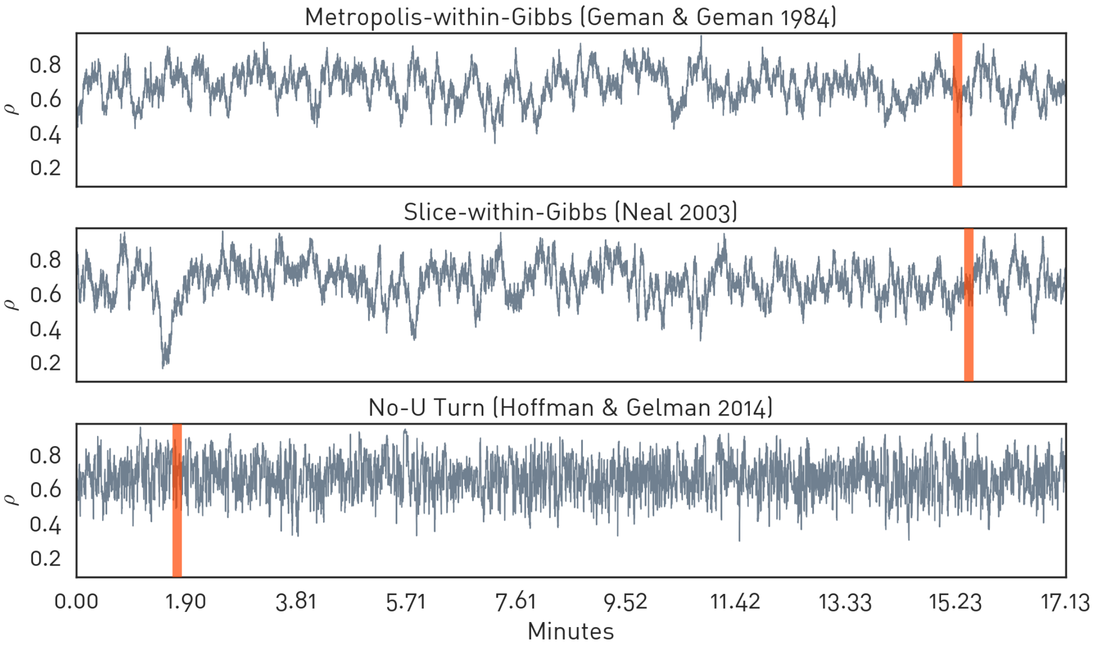

Figure 3: Traces for the spatial parameter in the hedonic model for house prices in Baltimore. Due to this case's low yields, the red vertical bar marks the expected time to take 100 effective samples instead of 1000.

In the second Baltimore demonstration, using the same spatial layout but an actual hedonic house price model with a subset of the covariates in Dubin (1992), statistics about effectiveness change substantially. The nominal size and throughput are similar to the results found for the first case. However, in the actual empirical model, the spatial autoregressive parameter is poorly behaved, as evidenced by the traces for ρ shown in Figure 3. The Metropolis sampler for ρ in this model attains an acceptance rate of $.233$, again nearly exactly the optimal rate suggested in Roberts et al. (1997). Despite this, the Metropolis trace exhibits significant random walk behavior, with the trace moving slowly around the conditional posterior. Similar strong random walk behavior is also shown by the Slice sampler. Thus, neither Gibbs strategy is very efficient. In contrast, NUTS proceeds unimpeded.

As shown in the second section of Table 1, the Metropolis sampler manages 115 independent draws in nearly $150,000$ iterations. While this means the samplers' yields are quite low, the higher throughput of the Gibbs samplers in this low-N is less relevant for the total effective size. NUTS performs well when compared to the other two methods, with a yield of about 11% and modest throughput. Most notably, the drop in yield between simulated data and real data for NUTS is not as dramatic as for the other two methods: its yield is halved in sampling parameters for the real data, whereas the other samplers drop two orders of magnitude in yield. Managing around $1,000$ independent samples in 17 minutes, NUTS generates 9 times more independent samples per second than Metropolis or Slice sampling. At their effective throughput, it would take Metropolis or Slice sampling over two and a half hours to draw $1,085$ effective samples, the same number that NUTS drew in twenty minutes.

4.1.2 Southern Counties Model

Moving on to the layout of the Southern U.S. counties, the issue of computing the Jacobian becomes more challenging. In contrast to the use of efficient sparse determinant routines used in many Maximum Likelihood, Slice-within-Gibbs, and Metropolis-within-Gibbs strategies, the NUTS implementations in `PyMC3` or `Stan` rely on off-the-shelf, dense-by-default automatic differentiation engines for generic inference. In our experiments, since yield and throughput are measured separately, we can separate the impact of scaling N on throughput independently from the differences in yield. We suggest that any throughput comparison should be understood to reflect the state of the art automatic HMC/NUTS available in common Bayesian computation packages, and not the maximum attainable computational efficiency of NUTS or HMC in general. However, as will be shown below, the NUTS implementation considered here simply does not handle these models efficiently.

The trace plots for the synthetic data (case three) are shown in Figure 4. In this instance, the NUTS trace appears rather thin when compared to Slice and Metropolis samples. This is because, for the $10,000$ iterations drawn using NUTS, the Metropolis-within-Gibbs implementation manages over a million samples and Slice draws around 350 thousand. Indeed, NUTS takes slightly more than 10 hours to conduct $10,000$ iterations. Sparse numerical operations in this case allow the Metropolis and Slice samplers to scale with N much more effectively than the NUTS implementation. This is critical, since NUTS, a much newer and sophisticated algorithm than either Slice or Metropolis-within-Gibbs, does not yet have implementations tuned to the special computational needs of the spatial model. In contrast, Slice or Metropolis samplers that use sparse matrix operations are common, so numerical efficiency trumps stochastic efficiency.

Examining the efficiency statistics presented in the lower half of Table 1, we see that the extreme

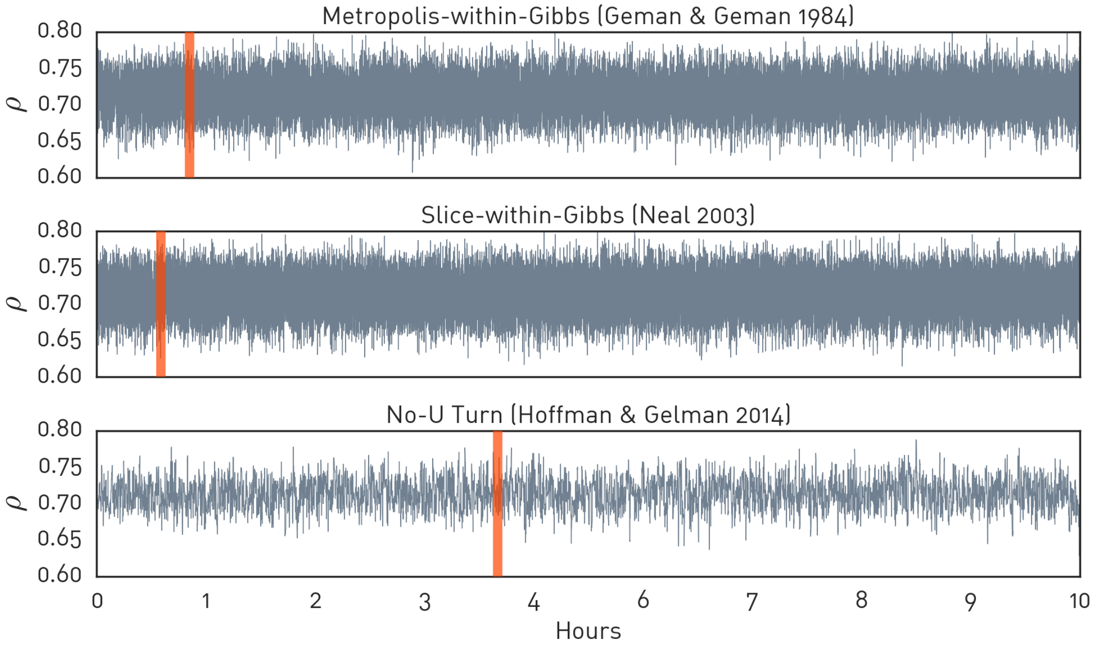

Figure 4: Traces for the spatial parameter in the synthetic model for Southern U.S. Counties. The red vertical bar marks the expected time to take 1000 effective samples.

difference in nominal throughput carries over into differences in effective throughput. While the NUTS algorithm has the same strong yield as in the small simulated data case, the massive relative throughput of the Metropolis sampler is able to decisively compensate for its lower yield in the larger data. The NUTS yield is around 16 times larger than the yield for Metropolis, but the Metropolis throughput is around 80 times larger than the NUTS throughput. This results in Metropolis drawing nearly four times the number of effectively independent samples, but they are drawn from over 80 times the iterations NUTS uses. Slice sampling achieves a modestly low yield of around 5%, similar to the result in the simulated small data example. However, this modest yield and throughput allows Slice sampling to outstrip both NUTS and Metropolis sampling, yielding a substantially larger effective sample size than either Metropolis or NUTS in the 10 hour sample run. For easier comparison, this means that $1,000$ effective samples are drawn in 40 minutes using Slice sampling. In contrast, drawing the equivalent of $1,000$ independent samples takes Metropolis an hour, and takes NUTS about 4 and a half hours.

For the colorectal cancer screening data in Southern counties, the corresponding trace is shown in Figure 5. Again, the thinness of the NUTS trace relative to the other two is clear. It takes NUTS nearly 15 hours to draw $10,000$ samples in the model for colorectal screening, whereas Slice sampling

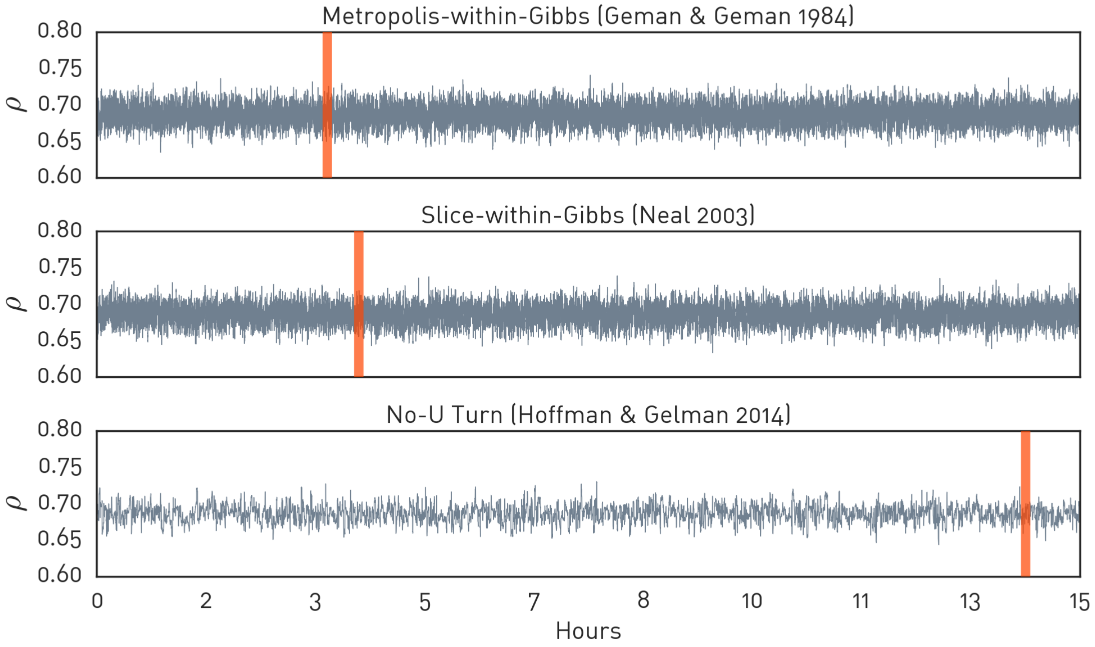

Figure 5: Traces for the spatial parameter in the colorectal cancer screening model for Southern U.S. Counties. The red vertical bar marks the expected time to take 1000 effective samples.

draws over forty times this amount, and Metropolis again outstrips all, drawing 1.8 million samples. But, as in the smaller-N scenario with real data, the Gibbs strategies exhibit strong random walk behavior in the traces. While NUTS is clearly hampered by its throughput, all of the sampler yields are again comparable to the small-N case with real data. In this case, Metropolis easily draws the largest effective sample in allotted time.

Finally, the change in sampler efficiency in simulated and empirical scenarios is similar regardless of N. In both scenario sizes, NUTS's yield is halved when moving from more well-behaved synthetic settings to a scenario more representative of that encountered in econometric practice. However, this is a smaller relative change in efficiency than that suffered by Slice or Metropolis sampling. Indeed, in the small-N scenario, Slice and Metropolis sampling drop two orders of magnitude in yield. In the larger case, they only drop a single order of magnitude in yield. Throughput is significantly affected by the change in problem size, however, with both Slice and Metropolis losing around a fifth of their throughput when moved to a dataset that is approximately seven times larger. While not necessarily desirable, this scales much better than the NUTS implementation considered here, which suffers a much larger drop in throughput in either case.

4.2 Impacts of Small Effective Size and Trace Autocorrelation

Given that the Slice and Metropolis-Hastings samplers may have quite low yield in empirical examples (where the data generating process is unknown), how does insufficient sample size even affect the conclusions econometricians draw from their models?[5] This most clearly connects with longstanding discussions in Bayesian statistics about "one long chain" versus "many short chains" (Gelman and Rubin, 1992b,a; Raftery and Lewis, 1992a,b; Kass et al., 1998). Recall, many spatial econometric papers use fixed nominal sample size stopping rules instead of fixed effective size/fixed posterior interval rules. Positive autocorrelation in sample traces also results in smaller effective sizes. So, a simulation that terminates on achieving a given nominal size may attain an unsatisfactory effective sample size because the effective sample size was never considered in the simulation stopping rule.

Critically, we do not suggest that econometricians simulate their chains for millions of iterations to achieve the same *effective* sample size as they might a *nominal* size common in the literature. However, if $10,000$ nominal samples contains an unacceptably low amount of effective samples, then stopping at $10,000$ (or $20,000$, etc.) is a premature termination: while point estimates may have converged, the posterior credible interval estimates may be too small relative to a correct long-run estimate or a more efficient sampler. The concern about insufficient effective size in fixed nominal size sampling runs is one of run length and mixing speed, not simulation convergence and ergodicity. If the chain is highly autocorrelated, early termination after the point estimates converge should result in smaller posterior intervals than anticipated, since a slowly-mixing chain fails to visit and return from the tails of the posterior distribution in that time. Instead, the slowly-mixing sampler spends most of its time near the posterior mode (Betancourt, 2017).

Therefore, we examine how the 95% highest posterior density interval (HPDI) evolves during the first $10,000$ iterations of the chains analyzed above. As suggested above, analyzing strongly-correlated series as if they were uncorrelated leads to overly-precise estimates of signal deviation. If this also affects short Markov chain Monte Carlo runs, then low effective sample sizes should indicate overly-precise posterior intervals.

[5] We are indebted in the next section focusing on the impact of ESS for inference to the work of a thourough, dedicated reviewer who raised cogent, consistent concerns with our initial exclusive focus on yield alone.

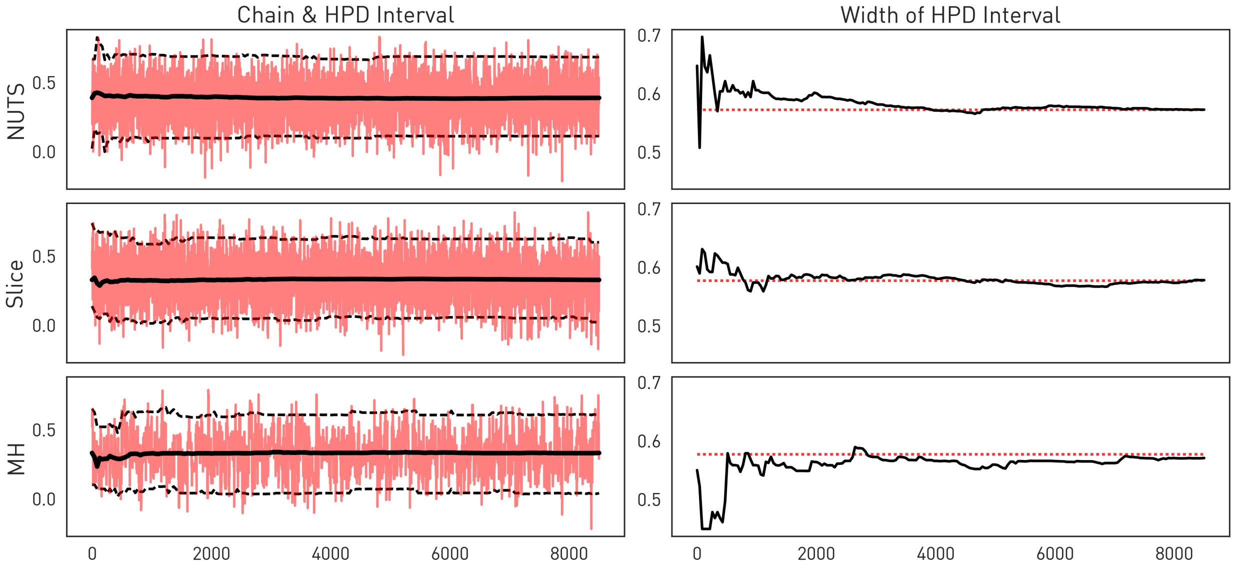

Figure 6: Traces (red), posterior medians (solid black), and 95% highest posterior density intervals (dashed black) for the first $10,000$ iterations of the Baltimore test case on left. On the right, the running 95% HPDI width is plotted in black, and the dotted red line is the long-run 95% HPDI width. Critically, since Slice and MH samplers take much more than $10,000$ iterations, this long-run estimate reflects the width of the 95% HPDI in total, not just the estimate at the end of $10,000$ iterations.

4.2.1 Baltimore Hedonic Model

For the well-behaved synthetic data, a plot of the first $10,000$ iterations of all samplers is shown in 6. According to its long-run yield reported in Table 1, the MH chain should have around 600 effective samples in its first $10,000$ iterations after convergence; NUTS and slice sampling should draw in excess of $2,000$. As shown in 6, only excessively early termination (within the first 500 iterations) would result in the HPDI being significantly understated. Both NUTS and Slice samplers tend to slightly overstate the HDPI, whereas the MH sampler tends to slightly understate it anywhere before the $10,000$th iteration. Regardless, the magnitude of this understatement is quite marginal. In addition, all samplers achieve similar long-run HPDI widths, as shown in Table 1. Thus, when N is small and trace autocorrelation is not severe, an effective size of 600 (and yield of 6%) does not understate the posterior credible interval.

In contrast, consider the highly-correlated parameter traces for the Baltimore hedonic house price model. Here, NUTS should draw around $1,000$ samples, whereas Slice and Metropolis samples should draw unacceptably small samples according to their yield. From the right side of Figure 7, termination any time in the first $10,000$ iterations would result in the MH sampler understating its long-run HPDI width by around 10%. In addition, termination of the Slice sampler before $5,000$ iterations would result

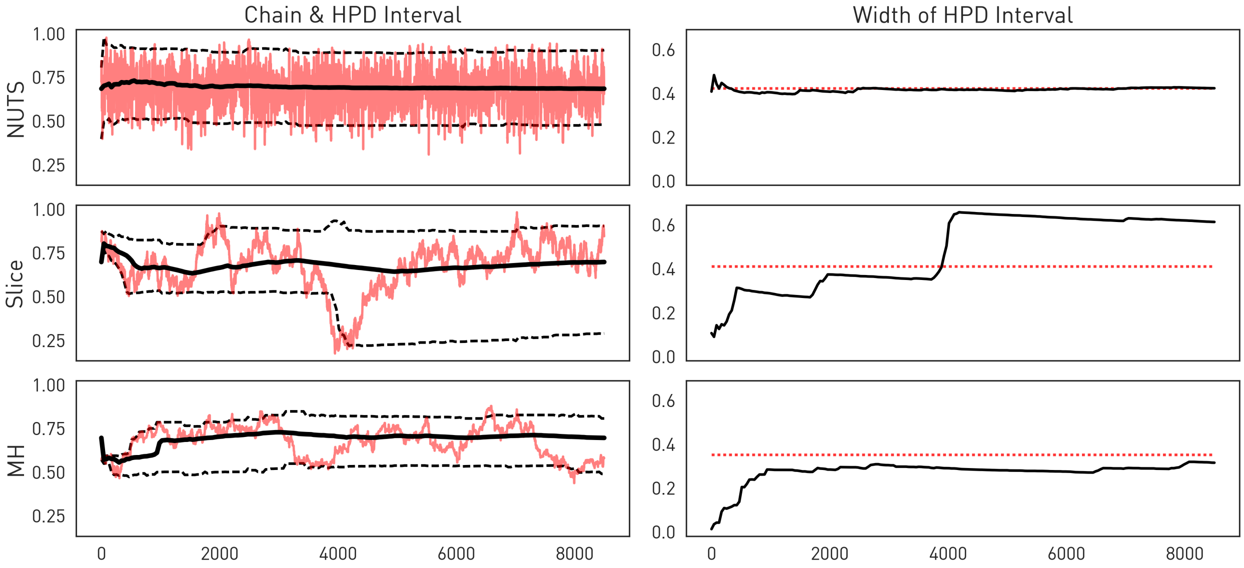

Figure 7: The first 10,000 iterations of the Baltimore hedonic house price model, symbolized in the same way as Fig. 6.

in significant understatement as well, but this understatement is unstable: after reaching a point value in the far tail of the distribution, the HPDI spends most of the remaining $50,000$ iterations decreasing to its long-run width. Finally, NUTS converges quickly to its long-run HDPI width. Critically, the MH long-run HPDI width, $.35$, is 12% smaller than the long-run width of the NUTS and Slice samplers, as shown in Table 1. Thus, any premature termination of the MH sampler in the first $10,000$ retained iterations will significantly understate an already understated long-run posterior interval width. Insufficient effective size in these first $10,000$ iterations would indicate that neither Slice or MH samplers have achieved a quality sample, even though their point estimates are largely stable.

Finally, for the large-N southern counties cases, a similar pattern manifests. Figure 8 provides the evolving traces, HPDIs, and HPDI widths. Here, the long-run MH width incorporates over $800,000$ iterations, and the long-run slice width incorporates over $350,000$, and NUTS exactly $10,000$. All samplers converge to their long-run HDPI widths in the first $10,000$ iterations, where NUTS draws over $2,000$ effective samples, Slice draws over 400, and MH over 100. Again, the size of the effective samples in the first $10,000$ iterations also provides an indication of how strongly a sampler hews to its long-run 95% HPDI interval. However, here, early termination of the NUTS sampler might result in an understated HPD interval. In addition, the NUTS HPDI long-run width is slightly smaller ($\sim 6\%$) than the Slice or MH long-run estimates. While it is possible this may increase if the simulation is allowed to run for longer,

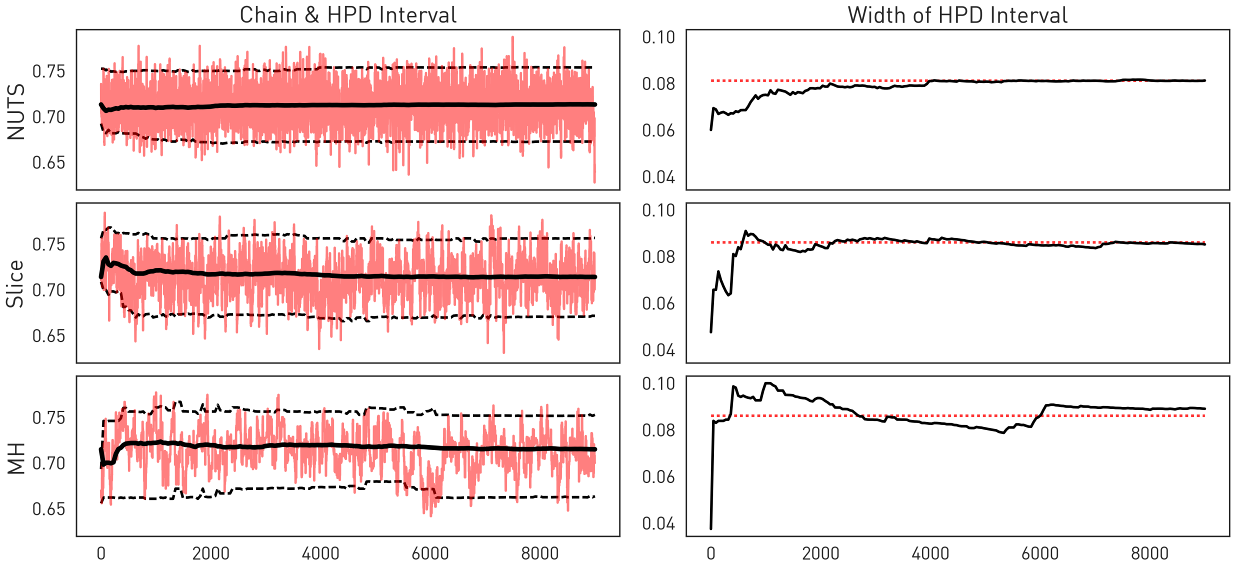

Figure 8: The first $10,000$ iterations of the southern counties test case, symbolized in the same way as Fig. 6.

the sheer computational inefficiency of NUTS in the big-N case makes it impractical for a researcher to routinely use long runs and compare their long-run with estimates.

Finally, the traces and HPDIs for colorectal cancer screenings in southern counties are shown in Figure 9. In the first $10,000$ iterations, NUTS would have an effective sample size of around $1,000$ samples according to its long-run yield, whereas Slice sampling generates an effective sample size of 75 and MH around 25 according to their long-run yields. In addition, all of the long-run HPDI widths are similar in this case, as shown in Table 1. But, this reflects the fact that Metropolis-Hastings and Slice samplers achieve significantly more iterations than NUTS, so their long-run HDPIs are substantially "longer" than NUTS. Understating that long-run HPDI width due to early termination is a significant concern in this case. While NUTS converges in the first $2,000$ iterations to its long-run HDPI, both Slice and Metropolis-Hastings samplers would tend to understate their long-run HPDI width. The Metropolis sampler requires nearly $7,500$ iterations before it gets within 10% of its long-run estimate and would significantly understate its long run HPDI width if stopped at any time before this.

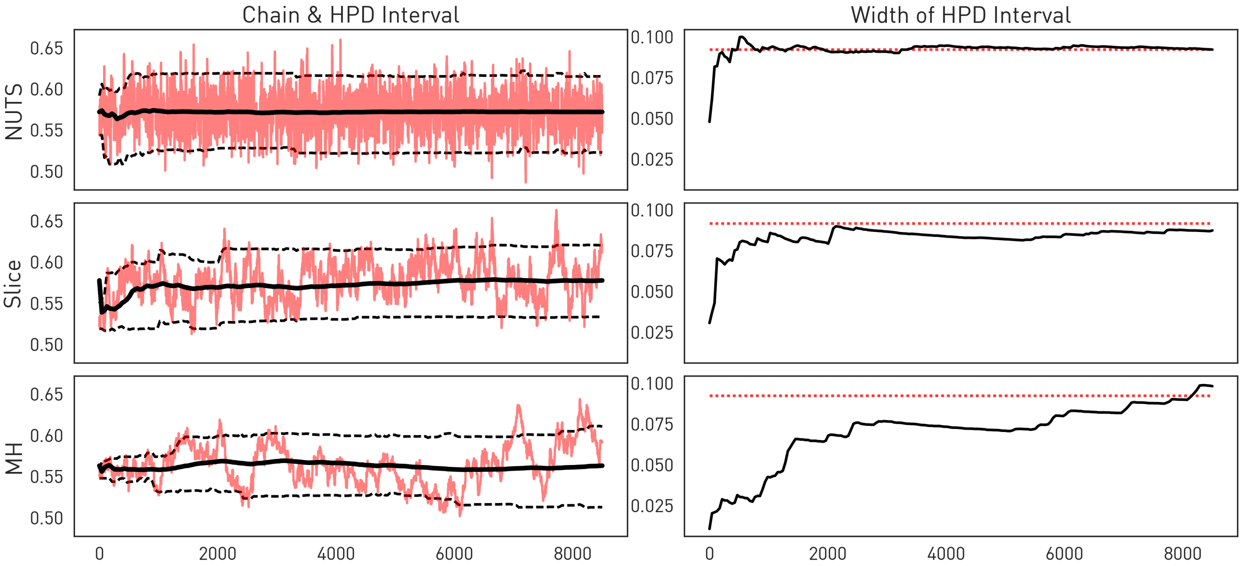

Figure 9: The first $10,000$ iterations of the southern counties colrectal cancer screening model, symbolized in the same way as Fig. 6.

5 Concluding Remarks

In terms of optimal stochastic and numerical efficiency, none of the three approaches dominate. Each have particular strong points, however. The Slice sampler seems to be the best default choice for drawing a fixed effective sample size in the smallest amount of time, as it has consistently acceptable effective throughput relative to the other samplers. In our study, in situations where the spatial autoregressive parameter exhibits significant random walk behavior, the loss of effective throughput places Slice sampling between NUTS and Metropolis sampling in terms of efficiency. However, when the parameter is well-behaved, slicing is extremely efficient, possibly surpassing NUTS. This is useful, since it is exceedingly simple to implement for univariate parameters. Slice sampling is not a new technique by any means and has been used extensively for spatial autoregressive parameters in Gaussian process models (Gelfand et al., 2003; Banerjee et al., 2014).

In contrast, NUTS, as a representative HMC sampler, shows itself as stochastically efficient, but is too numerically inefficient for spatial econometric models. Available NUTS implementations are too-low throughput to be competitive with other samplers reviewed. This is likely because the standard setup currently fails to take advantage of the special (Jacobian) structure of the spatial autoregressive model. If this were to be the case, considerable improvement in performance is likely, making the NUTS sampler a feasible strategy for spatial econometric modeling. For example, this may be accomplished by leveraging specialized libraries, such as `CHOLMOD` (Chen et al., 2008) for sparse Cholesky decomposition or the SuperLU library (Li, 2005) for sparse LU decomposition.

Our main finding pertains to the need for the adoption of better stopping rules in Bayesian spatial econometric practice and the routine use or reporting of a sample quality statistic, like effective sample size. This has been largely ignored in the Bayesian spatial econometric literature, where the main focus has been on fitting ever more complex spatial model specifications. Standard practice consists of using fixed nominal size stopping rules and the effective sample size or simulation yield is not reported. As our illustrations show, a fixed nominal size simulation may converge in its point estimates but understate the uncertainty about these estimates, especially when serial correlation in parameter draws is strong. This difference can be quite large, such as in the case of southern counties colorectal cancer screening example, where early termination would result in a significantly understated long-run HPDI width. In addition, the fact that a long-run HDPI may be understated relative to other sampling approaches suggests that basic research into better exact sampling methods is both useful and may provide insight into where current sampling methods under-perform.

Regardless, simulation quality should be quantified when studies are published. Effective sample size estimators are available in many different software packages and are relatively easy to interpret, understand, and present. In addition, efficient stopping rules based on effective sample size or posterior interval widths are also available. These stopping rules mitigate the potential surprise and early termination problems demonstrated in the chains considered in this paper. Finally, analysts and researchers should be sensitive to at least three measures of their simulations: *nominal throughput* (the number of raw samples generated), *throughput efficiency* (the number of effectively independent samples generated), and *yield* (the number of effectively independent samples generated per iteration). This work reinforces the insight that MCMC efficiency does not end at evaluating Metropolis rejection ratios or raw iterations per second, and that the search for *truly* effective general sampling techniques for Bayesian spatial econometric models remains an important area for future work.

Scenario	DGP	Sampler	Time (h:m:s)	Size	ESS	Tput	ETput	Yield	95% HPDI width
Baltimore	Synthetic	NUTS	00:19:58	10,000	2,339	8.35	1.95	23.39%	.5726
		Slice	00:19:58	58,801	15,017	49.04	12.52	25.54%	.5772
		Metropolis	00:19:58	139,401	8,592	116.26	7.17	6.16%	.5772
	Real	NUTS	00:17:07	10,000	1085	9.74	1.06	10.85%	.4232
		Slice	00:17:07	51,801	114	50.39	0.11	0.22%	.4114
		Metropolis	00:17:07	143,301	115	139.40	0.11	0.08%	.3528
Southern Counties	Synthetic	NUTS	10:02:30	10,000	2,043.04	0.28	0.06	20.43%	.0818
		Slice	10:02:30	355,601	15,001.75	9.82	0.41	4.22%	.0862
		Metropolis	10:02:30	839,001	10,462.40	23.17	0.29	1.25%	.0860
	Real	NUTS	14:47:05	10,000	1,091.08	0.19	0.02	10.91%	.0922
		Slice	14:47:05	469,701	3,679.69	8.81	0.07	0.78%	.0918
		Metropolis	14:47:05	1,800,901	4,511.23	33.84	0.08	0.25%	.0923

Table 1: Efficiency statistics for MCMC chains

References

Anselin, L. (1980). *Estimation methods for spatial autoregressive structures: a study in spatial econometrics*. PhD thesis, Cornell University.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Kluwer, Dordrecht.

Anselin, L. (2010). Thirty years of spatial econometrics. *Papers in Regional Science*, 89(1):325. 00188.

Anselin, L. and Rey, S. J. (2014). *Modern Spatial Econometrics in Practice, a Guide to GeoDa, GeoDaSpace, and PySAL*. GeoDa Press, Chicago, IL.

Baller, R. D., Anselin, L., Messner, S. F., Deane, G., and Hawkins, D. F. (2001). Structural covariates of U.S. county homicide rates: incorporating spatial effects. *Criminology*, 39(3):561588.

Baltagi, B. H., Egger, P. H., and Kesina, M. (2016). Bayesian Spatial Bivariate Panel Probit Estimation.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. CRC Press, Boca Raton, FL.

Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D., and Bengio, Y. (2010). Theano: A CPU and GPU math compiler in Python. In *Proceedings of the 9th Python in Science Conference*, page 17.

Besag, J. and Green, P. J. (1993). Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Socieity B*, 55(1):2537.

Betancourt, M. (2017). A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*.

Bivand, R., Hauke, J., and Kossowski, T. (2013). Computing the Jacobian in Gaussian spatial autoregressive models: an illustrated comparison of available methods. *Geographical Analysis*, 45(2):150179.

Bivand, R. and Piras, G. (2015). Comparing implementations of estimation methods for spatial econometrics. *Journal of Statistical Software*, 63(18):136.

Bivand, R. S., Gómez-Rubio, V., and Rue, H. (2014). Approximate Bayesian inference for spatial econometrics models. *Spatial Statistics*, 9:146165.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011). *Handbook of Markov Chain Monte Carlo*. CRC press, Boca Raton, FL.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo-Jiqiang, Li, P., and Riddell, A. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software*, In press.

Chen, Y., Davis, T. A., Hager, W. W., and Rajamanickam, S. (2008). Algorithm 887: CHOLMOD, supernodal sparse Cholesky factorization and update/downdate. *ACM Transactions on Mathematical Software (TOMS)*, 35(3):22.

Cliff, A. D. and Ord, J. K. (1981). *Spatial processes, models & applications*. Pion, London.

Cowles, M. K. and Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91(434):883904.

Diaconis, P., Holmes, S., and Neal, R. M. (2000). Analysis of a nonreversible Markov chain sampler. *Annals of Applied Probability*, 10(3):726752.

Dong, G. and Harris, R. (2015). Spatial autoregressive models for geographically hierarchical data structures. *Geographical Analysis*, 47(2):173191.

Dong, G. and Wu, W. (2016). Schools, land markets and spatial effects. *Land Use Policy*, 59:366374.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). Hybrid Monte Carlo. *Physics letters B*, 195(2):216222.

Dubin, R. A. (1992). Spatial autocorrelation and neighborhood quality. *Regional Science and Urban Economics*, 22(3):433452.

Finley, A. O., Banerjee, S., and Carlin, B. P. (2007). spBayes: an R package for univariate and multivariate hierarchical point-referenced spatial models. *Journal of Statistical Software*, 19(4):124.

Flegal, J. M. (2008). *Monte Carlo standard errors for Markov chain Monte Carlo*. PhD thesis, University of Minnesota.

Flegal, J. M., Haran, M., and Jones, G. L. (2008). Markov chain Monte Carlo: Can we trust the third significant figure? *Statistical Science*, 23(2):250260.

Fotheringham, A. S. and Brunsdon, C. (1999). Local forms of spatial analysis. *Geographical Analysis*, 31(4):340358.

Gelfand, A. E., Kim, H.-J., Sirmans, C. F., and Banerjee, S. (2003). Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, 98(462):387396.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515534.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014). *Bayesian data analysis*, volume 3. CRC Press, Boca Raton, FL.

Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, A., Roberts, G., and Gilks, W. (1996). Efficient Metropolis jumping rules. *Bayesian Statistics*, 5(599-608):42.

Gelman, A. and Rubin, D. B. (1992a). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457472.

Gelman, A. and Rubin, D. B. (1992b). A single series from the Gibbs sampler provides a false sense of security. *Bayesian Statistics*, 4:625631.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):721741.

Gong, L. and Flegal, J. M. (2016). A practical sequential stopping rule for high-dimensional Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics*, 25(3):684700.

Hepple, L. W. (1979). Bayesian analysis of the linear model with spatial dependence. In C.P.A. Bartels, R. K., editor, *Exploratory and explanatory statistical analysis of spatial data*, page 179199. Springer.

Hobert, J. P. and Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association*, 91(436):14611473.

Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):15931623.

Holloway, G., Shankar, B., and Rahmanb, S. (2002). Bayesian spatial probit estimation: a primer and an application to hyv rice adoption. *Agricultural Economics*, 27(3):383402.

Jensen, C. D., Lacombe, D. J., and McIntyre, S. G. (2013). A bayesian spatial econometric analysis of the 2010 uk general election. *Papers in Regional Science*, 92(3):651666.

Jones, G. L., Haran, M., Caffo, B. S., and Neath, R. (2006). Fixed-width output analysis for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 101(476):15371547.

Kass, R. E., Carlin, B. P., Gelman, A., and Neal, R. M. (1998). Markov chain monte carlo in practice: a roundtable discussion. *The American Statistician*, 52(2):93100.

Lacombe, D. J. and McIntyre, S. G. (2016). Local and global spatial effects in hierarchical models. *Applied Economics Letters*, 23(16):11681172.

Lee, D. (2013). CARBayes: an R package for Bayesian spatial modeling with conditional autoregressive priors. *Journal of Statistical Software*, 55(13):124.

LeSage, J. P. (1997). Bayesian estimation of spatial autoregressive models. *International Regional Science Review*, 20(1-2):113129.

LeSage, J. P. and Llano, C. (2013). A spatial interaction model with spatially structured origin and destination effects. *Journal of Geographical Systems*, 15(3):265289.

LeSage, J. P. and Pace, R. K. (2009). *Introduction to Spatial Econometrics*. CRC Press, Boca Raton, FL.

Li, X. S. (2005). An overview of SuperLU: Algorithms, implementation, and user interface. *ACM Transactions on Mathematical Software (TOMS)*, 31(3):302325.

Link, W. A. and Eaton, M. J. (2012). On thinning of chains in mcmc. *Methods in Ecology and Evolution*, 3(1):112115.

Lütkepohl, H. and Krätzig, M. (2004). *Applied time series econometrics*. Cambridge University Press.

Marshall, T. and Roberts, G. (2012). An adaptive approach to Langevin MCMC. *Statistical Computing*, 22(5):10411057.

Mobley, L., Kuo, T.-M., Urato, M., Boos, J., Lozano-Gracia, N., and Anselin, L. (2010a). Predictors of endoscopic colorectal cancer screening over time in 11 states. *Cancer Causes & Control*, 21(3):445461.

Mobley, L., Kuo, T.-M., Urato, M., and Subramanian, S. (2010b). Community contextual predictors of endoscopic colorectal cancer screening in the USA: spatial multilevel regression analysis. *International Journal of Health Geographics*, 9(1):1.

Neal, R. (1998). Suppressing random walks in Markov chain Monte Carlo using ordered overrelaxation. In Jordan, M., editor, *Learning in Graphical Models*, page 205228. Springer.

Neal, R. (2003). Slice sampling. *The Annals of Statistics*, 31(3):705767.

Neal, R. (2011). MCMC using Hamiltonian dynamics. In Brooks, S., Gelman, A., Jones, G., and Meng, X.-L., editors, *Handbook of Markov Chain Monte Carlo*, chapter 5. CRC Press.

Ohtsuka, Y., Kakamu, K., et al. (2015). Comparison of the sampling efficiency in spatial autoregressive model. *Open Journal of Statistics*, 5(01):10.

Pace, R. K. and Barry, R. (1997). Fast spatial estimation. *Applied Economics Letters*, 4(5):337341.

Pace, R. K. and LeSage, J. P. (2009). A sampling approach to estimate the log determinant used in spatial likelihood problems. *Journal of Geographical Systems*, 11(3):209225.

Parent, O. and LeSage, J. P. (2010). A spatial dynamic panel model with random effects applied to commuting times. *Transportation Research Part B: Methodological*, 44(5):633645.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News*, 6(1):711.

Polson, N. G. and Scott, J. G. (2012). On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4):887902.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raftery, A. E. and Lewis, S. (1992a). How many iterations in the Gibbs sampler. *Bayesian Statistics*, 4(2):763773.

Raftery, A. E. and Lewis, S. M. (1992b). [practical markov chain monte carlo]: comment: one long run with diagnostics: implementation strategies for markov chain monte carlo. *Statistical science*, 7(4):493497.

Rall, L. B. and Corliss, G. F. (1996). An introduction to automatic differentiation. In Berz, M., Bischof, C. H., and Corliss, G., editors, *Computational Differentiation: Techniques, Applications, and Tools*, volume 89 of *SIAM Proceedings in Applied Mathematics*, page 117, Santa Fe, NM. Society for Industrial and Applied Mathematics.

Ritter, C. and Tanner, M. A. (1992). Facilitating the Gibbs sampler: the Gibbs stopper and the griddy-Gibbs sampler. *Journal of the American Statistical Association*, 87(419):861868.

Robert, C. and Casella, G. (2011). A short history of Markov chain Monte Carlo: Subjective recollections from incomplete data. *Statistical Science*, 26(1):102115.

Roberts, G. O., Gelman, A., and Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7(1):110120.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, B*, 71(part 2):319392.

Salvatier, J., Wiecki, T. V., and Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2:e55.

Smirnov, O. and Anselin, L. (2001). Fast maximum likelihood estimation of very large spatial autoregressive models: a characteristic polynomial approach. *Computational Statistics & Data Analysis*, 35(3):301319.

Smith, T. E. and LeSage, J. P. (2004). A Bayesian probit model with spatial dependencies. *Advances in Econometrics*, 18:127160.

Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D. (2007). Openbugs user manual, version 3.0. 2. *MRC Biostatistics Unit, Cambridge*.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):17011728.

Wang, X. and Kockelman, K. M. (2009). Bayesian inference for ordered response data with a dynamic spatial-ordered probit model. *Journal of Regional Science*, 49(5):877913.

Wang, X. C., Kockelman, K. M., and Lemp, J. D. (2012). The dynamic spatial multinomial probit model: analysis of land use change using parcel-level data. *Journal of Transport Geography*, 24:7788.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, 41(3-4):434449.