Lecture 3: Bayesian estimation of spatial regression models

James P. LeSage
University of Toledo
Department of Economics
Toledo, OH 43606
jlesage@spatial-econometrics.com

October 24, 2005
1 Introduction to Bayesian variants of the spatial models

One might suppose that application of Bayesian estimation methods to SAR, SDM and SEM spatial regression models where the number of observations is very large would result in estimates nearly identical to those from maximum likelihood methods. This is a typical result when prior information is dominated by a large amount of sample information. Bayesian methods can however be used to relax the assumption of constant variance normal disturbances made by maximum likelihood methods, resulting in heteroscedastic Bayesian variants of the SAR, SDM and SEM models. In these models, the prior information exerts an impact, even in very large samples. In the next section, we discuss spatial heterogeneity to provide a motivation for relaxing the constant variance normality assumption in spatial relationships.

It is instructive to consider a Bayesian solution to the SAR estimation problem for the case of constant variance normal disturbances and diffuse priors, as this will facilitate our development of the heteroscedastic model. The likelihood for the SAR model can be written as in (1), where we express $A = I_n - \rho W$ for notational convenience.

$$L(\beta, \sigma, \rho, y, X) = (2\pi\sigma^2)^{-n/2}|A|\exp\left\{-\frac{1}{2\sigma^2}(Ay - X\beta)'(Ay - X\beta)\right\}$$ (1)

The likelihood can be combined with a prior for $\beta, \sigma$ taking the Jefferys' form: $p(\beta, \sigma|\rho) = \sigma^{-1}$. We rely on a general prior for $\rho$, that we denote as: $p(\rho)$. The product of the likelihood and prior represents the kernel posterior distribution for all parameters in our model via Bayes' theorem. We can represent this as:

$$p(\beta, \sigma, \rho|y, X, W) \propto |\sigma^{-n+1}|A|\exp\{-\frac{1}{2\sigma^2}(Ay - X\beta)'(Ay - X\beta)\}p(\rho)$$ (2)

We can treat $\sigma$ as a nuisance parameter and integrate this out using properties of the inverted gamma distribution, leading to:

$$p(\beta, \rho|y, X, W) \propto |A|\{(Ay - X\beta)'(Ay - X\beta)\}^{n/2}p(\rho)$$ (3)

$$= |A|\{(n-k)s^2(\rho) + (\beta - b(\rho))'X'X(\beta - b(\rho))\}^{-n/2}p(\rho)$$

Where:

$$b(\rho) = (X'X)^{-1}X'ay$$
$$s^2(\rho) = (Ay - Xb(\rho))'(Ay - Xb(\rho))/(n-k)$$
$$A = A(\rho) = (I_n - \rho W)$$ (4)

Conditional on $\rho$, the expression in (4) represents a multivariate Student-$t$ distribution that we can integrate with respect to $\beta$ leaving us with the marginal posterior distribution for $\rho$, shown in (5).

$$p(\rho|y, X, W) \propto |A(\rho)|(s^2(\rho))^{-(n-k)/2}p(\rho)$$ (5)
There is no analytical solution for the posterior expectation of $\rho$, which we would be interested in, so numerical integration methods would be required to find this expectation as well as the posterior variance of $\rho$. The integrals required are shown in (6).

$$E(\rho|y, X) = \bar{\rho} = \frac{\int \rho \cdot p(\rho|y, X, W)d\rho}{\int p(\rho|y, X, W)d\rho}$$

$$\text{var}(\rho|y, X) = \frac{\int [\rho - \bar{\rho}]^2 \cdot p(\rho|y, X)d\rho}{\int p(\rho|y, X)d\rho} \quad (6)$$

Using the direct sparse matrix approach or the Barry and Pace (1999) Monte Carlo estimator for the log determinant discussed in Lecture 1, to find $|A(\rho)|$ over a range of values from 0 to 1, facilitates numerical integration. This along with the vectorized expression for $s(\rho)^2 = \phi(\rho_i) = e_o' e_o - 2\rho_i e_o' e_o + \rho_i^2 e_o' e_o$, as a simple polynomial greatly simplifies numerical integration.

Given the posterior expectation, $\bar{\rho}$, the posterior expectation for the parameters $\beta$ can be computed using:

$$E(\beta|y, X, W) = (X'X)^{-1}X'(I_n - \bar{\rho}W)y \quad (7)$$

The variance-covariance matrix for the parameters $\beta$ is conditional on $\rho$, taking the form:

$$\text{var-cov}(\beta|\rho, y, X, W) = [(n - k)/(n - k - 2)]s^2(\rho)(X'X)^{-1} \quad (8)$$

which can be solved using univariate integration:

$$\text{var-cov}(\beta|y, X, W) = [(n - k)/(n - k - 2)]\{\int s^2(\rho)p(\rho|y, X, W)d\rho\}(X'X)^{-1} \quad (9)$$

Again, the Monte Carlo estimates for the log determinant as well as the vectorized expression for $s^2(\rho)$ facilitates computing the univariate integral:

$$E(s^2(\rho)|y, X, W) = \int s^2(\rho)p(\rho|y, X, W)d\rho \quad (10)$$

It is possible to solve large sample spatial problems using this approach in around twice the time required to solve for maximum likelihood estimates. However, this leaves us with an estimation approach that will tend to replicate maximum likelihood estimates in a computationally intensive way. The true benefits from applying a Bayesian methodology to spatial problems arise when we extend the conventional model to relax the constant variance normality assumptions placed on the disturbance process.

2 Spatial heterogeneity

The term spatial heterogeneity refers to variation in relationships over space. In the most general case we might expect a different relationship to hold for every point in space. Formally, we might write a linear relationship depicting this as:
\[ y_i = X_i \beta_i + \varepsilon_i \]  

(11)

Where \( i \) indexes observations collected at \( i = 1, \ldots, n \) points in space, \( X_i \) represents a \((1 \times k)\) vector of explanatory variables with an associated set of parameters \( \beta_i \), \( y_i \) is the dependent variable at observation (or location) \( i \) and \( \varepsilon_i \) denotes a stochastic disturbance in the linear relationship.

An important point is that we could not hope to estimate a set of \( n \) parameter vectors \( \beta_i \), as well as \( n \) noise variances \( \sigma_i \), given a sample of \( n \) data observations. We simply do not have enough sample data information with which to produce estimates for every point in space, a phenomena referred to as a “degrees of freedom” problem. To proceed with the analysis we need to provide a specification for variation over space. This specification must be parsimonious, that is, only a handful of parameters can be used in the specification. A large amount of spatial econometric research centers on alternative parsimonious specifications for modeling variation over space.

One can also view the specification task as one of placing restrictions on the nature of variation in the relationship over space. For example, suppose we classified our spatial observations into urban and rural regions. We could then restrict our analysis to two relationships, one homogeneous across all urban observational units and another for the rural units. This raises a number of questions: 1) are two relations consistent with the data, or is there evidence to suggest more than two?, 2) is there a trade-off between efficiency in the estimates and the number of restrictions we use?, 3) are the estimates biased if the restrictions are inconsistent with the sample data information?.

One of the compelling motivations for the use of Bayesian methods in spatial econometrics is their ability to impose restrictions that are stochastic rather than exact in nature. Bayesian methods allow us to impose restrictions with varying amounts of prior uncertainty. In the limit, as we impose a restriction with a great deal of certainty, the restriction becomes exact. Carrying out our econometric analysis with varying amounts of prior uncertainty regarding a restriction allows us to provide a continuous mapping of the restriction’s impact on the estimation outcomes.

3 Bayesian heteroscedastic spatial models

We introduce a more general version of the SAR, SDM and SEM models that allows for non-constant variance across space as well as outliers. When dealing with spatial datasets one can encounter what have become known as “enclave effects”, where a particular region does not follow the same relationship as the majority of spatial observations. For example, all counties in a single state might represent aberrant observations that differ from those in all other counties. This will lead to fat-tailed errors that are not normal, but more likely to follow a Student-\( \text{t} \) distribution.

This extended version of the SAR, SDM and SEM models involves introduction of non-constant variance to accommodate spatial heterogeneity and outliers that arise in applied practice. Here we can follow LeSage (1997, 2000) and introduce a set of variance scalars \((v_1, v_2, \ldots, v_n)\), as unknown parameters that need to be estimated. This allows us to assume \( \varepsilon \sim N(0, \sigma^2 V) \), where \( V = \text{diag}(v_1, v_2, \ldots, v_n) \). The prior distribution for the \( v_i \) terms takes
the form of an independent $\chi^2(r)/r$ distribution. Recall that the $\chi^2$ distribution is a single parameter distribution, where we have represented this parameter as $r$. This allows us to estimate the additional $n$ parameters $v_i$ in the model by adding the single parameter $r$ to our estimation procedure.

This type of prior was used in Geweke (1993) to model heteroscedasticity and outliers in the context of linear regression. The specifics regarding the prior assigned to the $v_i$ terms can be motivated by considering that the mean equals unity and the variance of the prior is $2/r$. This implies that as $r$ becomes very large, the terms $v_i$ will all approach unity, resulting in $V = I_n$, the traditional assumption of constant variance across space. On the other hand, small values of $r$ lead to a skewed distribution that permits large values of $v_i$ that deviate greatly from the prior mean of unity. The role of these large $v_i$ values is to accommodate outliers or observations containing large variances by downweighting these observations. Note that $\varepsilon \sim N(0, \sigma^2 V)$, with $V$ diagonal implies a generalized least-squares (GLS) correction to the vector $y$ and explanatory variables matrix $X$. The GLS correction involves dividing through by $\sqrt{v_i}$, which leads to large $v_i$ values functioning to downweight these observations. Even in large samples, this prior will exert an impact on the estimation outcome.

A formal statement of the Bayesian heteroscedastic SAR model is shown in (12), where we have added a normal-gamma conjugate prior for $\beta$ and $\sigma$, and a uniform prior for $\rho$ in addition to the chi-squared prior for the terms in $V$. The prior distributions are indicated using $\pi$.

\[
\begin{align*}
y &= \rho W y + X \beta + \varepsilon \\
\varepsilon &\sim N(0, \sigma^2 V) \quad V = \text{diag}(v_1, \ldots, v_n) \\
\pi(\beta) &\sim N(c, T) \\
\pi(r/v_i) &\sim \text{IID } \chi^2(r) \\
\pi(1/\sigma^2) &\sim \Gamma(d, \nu) \\
\pi(\rho) &\sim U[0, 1]
\end{align*}
\] (12)

In the case of very large samples involving upwards of 10,000 observations, the normal-gamma priors for $\beta, \sigma$ should exert relatively little influence. Setting $c$ to zero and $T$ to a very large number results in a diffuse prior for $\beta$. Diffuse settings for $\sigma$ involve setting $d = 0, \nu = 0$. For completeness, we develop the results for the case of a normal-gamma prior on $\beta, \sigma$.

In contrast to the case of the priors on $\beta, \sigma$, assigning an informative prior to the parameter $\rho$ associated with spatial dependence should exert an impact on the estimation outcomes even in large samples. This is due the important role played by spatial dependence in these models. In typical applications where the magnitude and significance of $\rho$ is a subject of interest, a diffuse prior would be used. It is however possible to rely on an informative prior for this parameter.
4 Estimation of Bayesian spatial models

An unfortunate complication that arises with this extension is that the addition of the chi-squared prior greatly complicates the posterior distribution, ruling out the simple univariate numerical integration approach outlined in section 2. Assume for the moment, diffuse priors for $\beta, \sigma$. A key insight is that if we knew $V$, this problem would look like a GLS version of the previous problem from section 2. That is, conditional on $V$, we would arrive at similar expressions as in section 2, where the $y$ and $X$ are transformed by dividing through by: $\sqrt{\text{diag}(V)}$. We rely on a Markov Chain Monte Carlo (MCMC) estimation method that exploits this fact.

MCMC is based on the idea that a large sample from the posterior distribution of our parameters can be used in place of an analytical solution where this is difficult or impossible. We designate the posterior using $p(\theta | D)$, where $\theta$ represents the parameters and $D$ the sample data. If the sample from $p(\theta | D)$ were large enough, we could approximate the form of the posterior density using kernel density estimators or histograms, eliminating the need to know the precise analytical form of this complicated density. Simple statistics can be used to construct means and variances based on the sample from the posterior.

The parameters $\beta, V$ and $\sigma$ in the heteroscedastic SAR model can be estimated by drawing sequentially from the conditional distributions of these parameters, a process known as Gibbs sampling because of its origins in image analysis, (Geman and Geman (1984)). It is also labeled “alternating conditional sampling”, which seems a more accurate description. To illustrate how this works, let $\theta = (\theta_1, \theta_2)$, represent a parameter vector and $p(\theta)$ denote the prior, with $L(\theta | y, X, W)$ denoting the likelihood. This results in a posterior distribution $p(\theta | D) = c \cdot p(\theta)L(\theta | y, X, W)$, with $c$ a normalizing constant. Consider the case where $p(\theta | D)$ is difficult to work with, but a partition of the parameters into two sets $\theta_1, \theta_2$ is easier to handle. Given an initial estimate for $\theta_1$, which we label $\hat{\theta}_1$, suppose we could easily estimate $\theta_2$ conditional on $\theta_1$ using $p(\theta_2 | D, \hat{\theta}_1)$. Denote the estimate, $\hat{\theta}_2$ derived by using the posterior mean or mode of $p(\theta_2 | D, \hat{\theta}_1)$. Assume further that we are now able to easily construct a new estimate of $\theta_1$ based on the conditional distribution $p(\theta_1 | D, \hat{\theta}_2)$. This new estimate for $\theta_1$ can be used to construct another value for $\theta_2$, and so on. On each pass through the sequence of sampling from the two conditional distributions for $\theta_1, \theta_2$, we collect the parameter draws which are used to construct a joint posterior distribution for the parameters in our model. Gelfand and Smith (1990) demonstrate that sampling from the sequence of complete conditional distributions for all parameters in the model produces a set of estimates that converge in the limit to the true (joint) posterior distribution of the parameters. That is, despite the use of conditional distributions in our sampling scheme, a large sample of the draws can be used to produce valid posterior inferences regarding the joint posterior mean and moments of the parameters.

4.1 Why it works

You might wonder why this works — after all it seems counter-intuitive. Alternating conditional sampling has its roots in early work of Metropolis, et al. (1953), who showed that one could construct a Markov chain stochastic process for $(\theta_t, t \geq 0)$ that unfolds over time such that: 1) it has the same state space (set of possible values) as $\theta$, 2) it is easy to sim-
ulate, and 3) the equilibrium or stationary distribution which we use to draw samples is \( p(\theta|D) \) after the Markov chain has been run for a long enough time. Given this result, we can construct and run a Markov chain for a very large number of iterations to produce a sample of \((\theta_t, t = 1, \ldots)\) from the posterior distribution and use simple descriptive statistics to examine any features of the posterior in which we are interested.

The most widely used approach to MCMC is due to Hastings (1970) which generalizes the method of Metropolis et al. (1953). Hastings (1970) suggests that given an initial value \( \theta_0 \) we can construct a chain by recognizing that any Markov chain that has found its way to a state \( \theta_t \) can be completely characterized by the probability distribution for time \( t + 1 \).

His algorithm relies on a proposal or candidate distribution, \( f(\theta|\theta_t) \) for time \( t + 1 \), given that we have \( \theta_t \). A candidate point \( \theta^* \) is sampled from the proposal distribution and:

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

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

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

In other words, we can view the Hastings algorithm as indicating that we should toss a Bernoulli coin with probability \( \psi_H \) of heads and make a move to \( \theta_{t+1} = \theta^* \) if we see a heads, otherwise set \( \theta_{t+1} = \theta_t \). Hastings demonstrates that this approach to sampling represents a Markov chain with the correct equilibrium distribution capable of producing samples from the posterior \( p(\theta|D) \) we are interested in.

The Gibbs or alternating conditional sampling approach described above, represents a special case of the Metropolis-Hastings algorithm, where every draw is accepted, see Gelman, Carlin, Stern and Rubin (1995, p. 328).

### 4.2 Conditional distributions

To implement this estimation method, we need to determine the conditional distributions for each parameter in our Bayesian heteroscedastic SAR model. The conditional distribution for \( \beta \) follows from the insight that given \( V \), we can rely on standard Bayesian GLS regression results to show that:

\[
\begin{align*}
   p(\beta|\rho, \sigma, V) & \sim N(\bar{\beta}, \sigma^2B) \\
   \bar{\beta} & = (X'V^{-1}X + \sigma^2T^{-1})^{-1}(X'V^{-1}(I_n - \rho W)y + \sigma^2T^{-1}e) \\
   B & = \sigma^2(X'V^{-1}X + \sigma^2T^{-1})^{-1}
\end{align*}
\]

We see that the conditional for \( \beta \) is a multinormal distribution from which it is easy to sample a vector \( \beta \).

The conditional distribution for \( \sigma \) given the other parameters, takes the form (see Gelman, Carlin, Stern and Rubin, 1995):

\[
\begin{align*}
p(\sigma^2|\beta, \rho, V) & \propto (\sigma^2)^{-(n+d+1)/2} \exp \left[ -e'V^{-1}e + \frac{2\nu}{2\sigma^2} \right] \\
e & = (I_n - \rho W)y - X\beta
\end{align*}
\]
which is proportional to an inverse gamma distribution with parameters \((n/2) + d\) and \(e'Ve^{-1}e + 2\nu\). Again, this would be an easy distribution from which to sample a scalar value for \(\sigma\).

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

\[
p\left(\frac{e_i^2 + r}{v_i} \mid \beta, \rho, \sigma^2, v_{-i}\right) \sim \chi^2(r + 1)
\]

where \(v_{-i} = (v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_n)\) for each \(i\), and \(e\) is as defined in (16). Again, this represents a known distribution from which it is easy to construct a scalar draw.

Finally, the conditional posterior distribution of \(\rho\) takes the form:

\[
p(\rho \mid \beta, \sigma, V) \propto |A(\rho)| (s^2(\rho))^{-(n-k)/2} p(\rho)
\]

\[
s^2(\rho) = (Ay - Xb(\rho))^\prime V^{-1} (Ay - Xb(\rho)) / (n - k)
\]

A problem arises here in that this distribution is not one for which established algorithms exist to produce random draws. There are however ways to sample from an arbitrary distribution such as this, for example the Metropolis-Hastings algorithms could be used with a suitable proposal density. This approach that relies on Metropolis-Hastings sampling for the parameter \(\rho\) within a sequence of Gibbs sampling steps to obtain \(\beta, \sigma\) and \(V\), represents a procedure that is often labeled “Metropolis within Gibbs sampling” (Gelman, Carlin, Stern and Rubin, 1995).

LeSage (2000) suggests a normal or Student-\(t\) distribution as a proposal for the Metropolis-Hastings step to obtain \(\rho\) and develops algorithms implementing this approach for the Bayesian heteroscedastic SAR, and other spatial models. LeSage (1997) achieves the same goal using a “ratio-of-uniforms” method rather than the Metropolis-Hastings approach. In fact, if you think about it, computational generation of random deviates from a vast array of statistical distributions is accomplished beginning with computer-generated uniform random deviates, and there is an extensive literature on this topic.

### 4.3 The MCMC sampler

By way of summary, an MCMC estimation scheme involves starting with arbitrary initial values for the parameters which we denote \(\beta^0, \sigma^0, V^0, \rho^0\). We then sample sequentially from the following set of conditional distributions for the parameters in our model.

1. \(p(\beta \mid \sigma^0, V^0, \rho^0)\), which is a multinormal distribution with mean and variance defined in (14) and (15). This updated value for the parameter vector \(\beta\) we label \(\beta^1\).

2. \(p(\sigma \mid \beta^1, V^0, \rho^0)\), which is chi-squared distributed with \(n + 2d\) degrees of freedom as shown in (16). Note that we rely on the updated value of the parameter vector \(\beta = \beta^1\) when evaluating this conditional density. We label the updated parameter \(\sigma = \sigma^1\) and note that we will continue to employ the updated values of previously sampled parameters when evaluating the next conditional densities in the sequence.
3. \( p(v_i|\beta^1, \sigma^1, v_{-i}, \rho^0) \) which can be obtained from the chi-squared distribution shown in (18). Note that this draw can be accomplished as a vector, providing greater speed.

4. \( p(\rho|\beta^1, \sigma^1, V^1) \), which we could sample using a metropolis step based on a normal or Student-\( t \) candidate distribution. We can also constrain \( \rho \) to an interval such as (0,1) using rejection sampling. This simply means that we reject values of \( \rho \) outside this interval. Note also that it is easy to implement a normal or some alternative prior distribution for this parameter.

We now return to step 1) employing the updated parameter values in place of the initial values \( \beta^0, \sigma^0, V^0, \rho^0 \). On each pass through the sequence we collect the parameter draws which are used to construct a joint posterior distribution for the parameters in our model. As already noted, Gelfand and Smith (1990) demonstrate that MCMC sampling from the sequence of complete conditional distributions for all parameters in the model produces a set of estimates that converge in the limit to the true (joint) posterior distribution of the parameters. Another point to note is that the parameter draws can be used to test hypotheses regarding any function of interest involving the parameters.

4.4 A recent innovation

Recently, I have discovered a fast and more accurate approach than Metropolis-Hasting to obtain draws for \( \rho \) during MCMC sampling of the spatial SAR, SDM and SEM models. Given the computational power of today’s computers, one can in fact produce a draw from the conditional distribution for \( \rho \) in these models using univariate numerical integration on each pass through the sampler. A few years back, this would have been unthinkable, as numerical integration represented one of the more computationally demanding tasks.

Here is the approach based on numerical integration of the conditional posterior of \( \rho \) that I now recommend for tackling estimation of these problems. Use logs to transform the conditional posterior in (19), and the Barry and Pace (1999) Monte Carlo estimator for the log determinant of today’s computers, one can in fact produce a draw from the conditional distribution for \( \rho \) in these models using univariate numerical integration on each pass through the sampler. A few years back, this would have been unthinkable, as numerical integration represented one of the more computationally demanding tasks.

Here is the approach based on numerical integration of the conditional posterior of \( \rho \) that I now recommend for tackling estimation of these problems. Use logs to transform the conditional posterior in (19), and the Barry and Pace (1999) Monte Carlo estimator for the log determinant in (19), along with the vectorized expression for \( s(\rho)^2 = \phi(\rho_i) = e'_o e_o - 2 \rho i e'_o e_o + \rho_i^2 e'_o e_d \). This produces a simple numerical integration problem that can be solved rapidly using Simpson’s rule. We arrive at the entire conditional distribution using this numerical integration approach, and then produce a draw from this distribution using “inversion”. Keep in mind that on the next pass through the MCMC sampler, we need to integrate the conditional posterior again. This is because the distribution is conditional on the changing values for the other parameters \( v_i, \beta, \sigma \) in the model, which obviously produce an altered expression for \( s^2 \) in the conditional distribution for \( \rho \).

5 Applied examples

To illustrate the heteroscedastic Bayesian SAR model, we carry out an experiment where we generate our own data based on a small 49 observation spatial dataset from Anselin (1988). The latitude-longitude coordinates from 49 neighbors in Columbus, Ohio were used to construct a spatial weight matrix, \( W \). The SAR model was generated using:

\[
y = (I_n - \rho W)^{-1} X \beta + (I_n - \rho W)^{-1} \varepsilon
\]  

(20)
where $\beta = (1, 1, 1)'$, $\rho = 0.7$, and the matrix $X \sim N(0, 1)$. To create heteroscedasticity in the disturbance process, an $n$-vector of $N(0, 0.1)$ disturbances was generated and observations 11 to 20 were multiplied by 5, creating an inflated variance for part of the sample. An important point here is that heteroscedasticity or aberrant observations will tend to exert an influence on neighboring observations through $(I_n - \rho W)^{-1}$ (see the matrix expressions (4) and (5) from Lecture #1).

A set of 100 data vectors $y$ were generated, with the matrix $X$ held fixed, but the disturbances generated randomly on each trial. A set of 100 estimates were carried out using: 1) the maximum likelihood SAR model, 2) the Bayesian heteroscedastic model with $r = 200$, a homoscedastic prior, and 3) the Bayesian model with $r = 4$, a heteroscedastic prior. This last model should produce the best estimates. The Bayesian model based on the homoscedastic prior should produce estimates that are more similar to those from maximum likelihood.

The MATLAB program to carry out this experiment is shown below.

```matlab
% example2.m file
% we generate our own heterosecastic SAR model
load anselin.dat; % 49 observations on Columbus, Ohio neighbors
latt = anselin(:,4); long = anselin(:,5);
% create W-matrix using Anselin's neigbhorhood crime data set
[junk W junk] = xy2cont(latt,long);
[n junk] = size(W); k = 3; IN = eye(n);
sige = 0.1; rho = 0.7; % true values for sige and rho
x = randn(n,k); % generate random normal X-matrix
beta = ones(k,1); % true parameter values
ndraw = 2500; % # of draws to carry out
nomit = 500; % # of draws to exclude for burn-in
% generate fixed part of the SAR model
y = inv(IN-rho*W)*x*beta;

niter = 100; % the # of experiments
maxl = zeros(niter,4); % storage for ML estimates
homo = zeros(niter,4); % storage for Bayes homoscedastic estimates
hetr = zeros(niter,4); % storage for Bayes heteroscedastic estimates
vout = zeros(niter,n); % storage for vi-estimates
tic; % turn on the timer
for i=1:niter; % do niter replications
    randn('seed',i); % control random number generation
evec = randn(n,1)*sqrt(sige); % constant variance
evec(11:20,1) = 5*evec(11:20,1); % add non-constant variance
y = y + inv(IN-rho*W)*evec; % add disturbances
% estimate maximum likelihood model
info.rmin = 0; % limits on rho
info.rmax = 1;
resl = sar(y,x,W,info);
maxl(i,:) = [resl.beta' resl.rho]; % save estimates
% MCMC sampling estimates
prior.rval = 200; % homoscedastic prior
% this is the c-mex version, sar_g is the matlab version
results2 = sar_gc(y,x,W,ndraw,nomit,prior);
homo(i,:) = [mean(results2.bdraw) mean(results2.pdraw)];
prior.rval = 4; % heteroscedastic prior
results3 = sar_gc(y,x,W,ndraw,nomit,prior);
hetr(i,:) = [mean(results3.bdraw) mean(results3.pdraw)];
```
vout(i,:) = results3.vmean'; % save vi-estimates
end;
toc; % report the time needed

plot(mean(vout)); % plot mean over niter experiments of the estimates for vi
xlabel('observations'); ylabel('posterior mean of vi-draws over 100 experiments');

out = [mean(maxl)' std(maxl)' mean(homo)' std(homo)' mean(hetr)' std(hetr)'];
fmt.cnames = strvcat('maxl','maxl std','homo','homo std','hetr','hetr std');
fmt.rnames = strvcat('variables','b1(=1)','b2(=1)','b3(=1)','rho(=0.7)');
mprint(out,fmt); % mprint() is a spatial econometrics toolbox function
% for pretty printing matrices

[h1 f1 y1] = pltdens(maxl(:,1));
h2 f2 y2] = pltdens(hetr(:,1)); % pltdens() is a spatial econometrics toolbox function
subplot(2,1,1),
plot(y1,f1,'o',y2,f2,'+'); % legend('max like','Bayes hetero');
xlabel('eta_1 values');
ylabel('Distribution of 100 experiments');

Before discussing the results, note that it took 371 seconds to produce maximum likelihood estimates as well as 2500 draws for both the homoscedastic and heteroscedastic Bayesian models over all 100 experiments. The number of draws carried out over the 100 experiments was 50,000, or 138 draws per second — truly amazing given that we rely on numerical integration on every pass through the MCMC sampler. These times were for an Athlon 1200 Mhz. computer with 768 megabytes of DDR memory, using c-mex routines that interface to MATLAB. Other timing results are shown below for a slower computer where the size of the estimation problem was varied from 3,000 observations up to 35,702.

650 Mhz Pentium III laptop
sar model: y = p*W*y + X*b + e
==================================================================
n = 3,000 observations, k=7 variables
  max lik  sar_gc
  total time in secs =  5.75 21.5310
  time for sampling (1000 draws) 17.9960
==================================================================
n = 9,000 observations, k=7 variables
  max lik  sar_gc
  total time in secs = 16.2440 94.6260
  time for sampling (1000 draws) 84.0910
==================================================================
n = 18,000 observations, k=7 variables
  max lik  sar_gc
  total time in secs = 31.5250 192.2560
  time for sampling (1000 draws) 169.7940

10
The means and standard deviations for the distribution of estimation outcomes based on 100 experiments are shown below, as they were printed by the program. They point to somewhat improved accuracy of the Bayesian heteroscedastic model versus the other two. As we expected, the Bayesian homoscedastic model produced estimates very near those from maximum likelihood. This model has essentially diffuse priors on all parameters, so we should expect this result. Another point is that the Bayesian heteroscedastic estimates produced a consistently smaller dispersion in the estimation outcomes across all of the parameters. In least-squares, heteroscedasticity creates an efficiency problem, but not bias, so perhaps this is an expected result.

<table>
<thead>
<tr>
<th>variables</th>
<th>maxl</th>
<th>maxl std</th>
<th>homo</th>
<th>homo std</th>
<th>hetr</th>
<th>hetr std</th>
</tr>
</thead>
<tbody>
<tr>
<td>b1 (=1)</td>
<td>1.2595</td>
<td>0.1613</td>
<td>1.2486</td>
<td>0.1533</td>
<td>1.0053</td>
<td>0.0687</td>
</tr>
<tr>
<td>b2 (=1)</td>
<td>0.9880</td>
<td>0.1155</td>
<td>0.9860</td>
<td>0.1140</td>
<td>0.9495</td>
<td>0.0819</td>
</tr>
<tr>
<td>b3 (=1)</td>
<td>1.2302</td>
<td>0.1773</td>
<td>1.2137</td>
<td>0.1668</td>
<td>0.9948</td>
<td>0.0640</td>
</tr>
<tr>
<td>rho (=0.7)</td>
<td>0.6456</td>
<td>0.1106</td>
<td>0.6422</td>
<td>0.0924</td>
<td>0.6761</td>
<td>0.0729</td>
</tr>
</tbody>
</table>

Plots of the distribution of estimate outcomes based on a kernel density estimate proves very enlightening. Figure 1 shows these distributions for both the maximum likelihood and Bayesian heteroscedastic models. The distribution of outcomes for the $\beta_1$ estimates produced by the Bayesian model is clearly superior to the distribution of maximum likelihood estimates. A similar result (not shown) held for the $\beta_3$ parameter. Superiority in the case of the distribution of $\beta_2$ estimates is a little less clear.
Figure 1: Distribution of 100 estimate outcomes for $\beta_1, \beta_2, \rho$

The distribution of outcomes for $\rho$ are also shown in Figure 1 where we see a very long left-tail in the maximum likelihood and homoscedastic Bayesian model outcomes, pointing
to low levels of spatial dependence extending to 0.3. In contrast, the Bayesian heteroscedastic estimates are all above 0.45, and exhibit smaller dispersion in the right-tail as well. The modes of the two non-heteroscedastic distributions confirm the downward bias in the means reported in the printed results. The impact of non-constant variance appears to be a downward bias in the maximum likelihood estimate of spatial dependence measured by $\rho$.

Finally, the $v_i$ estimates for individual observations can be used to detect regions in the spatial sample where aberrant observations or non-constant variance exist. The means from the Bayesian heteroscedastic model $v_i$ estimates over the 100 experiments are shown in Figure 2. To conserve on memory, the MATLAB functions `sar_g`, `sar_ge` return only the mean of the $v_i$ draws. Returning all draws would require storage of an $n$ by ndraw matrix where ndraw is usually in the thousands. The pattern of higher variance over observations 11 to 20 was clearly captured by the heteroscedastic model $v_i$ estimates.

References


