Rose-Hulman Undergraduate Mathematics Journal

Bayesian estimation in autoregressive models using reversible jump Markov chain Monte Carlo

Nathan Rogers[a]

Volume 16, No. 2, Fall 2015

Sponsored by

Rose-Hulman Institute of Technology
Department of Mathematics
Terre Haute, IN 47803
Email: mathjournal@rose-hulman.edu
http://www.rose-hulman.edu/mathjournal

[a] University of Illinois at Chicago

Rose-Hulman Undergraduate Mathematics Journal
Volume 16, No. 2, Fall 2015

Bayesian estimation in autoregressive models using reversible jump Markov chain Monte Carlo

Nathan Rogers

Abstract. In most applications, there is uncertainty about the statistical model to be considered. In this paper, we consider a particular class of autoregressive time series models where the order of the model—which determines the dimension of parameter—is uncertain. A common approach for model selection is to balance model fit with model complexity using, say, an AIC criterion. However, such an approach provides no meaningful measure of uncertainty about the selected model. A Bayesian approach, on the other hand, which treats the model and model parameters as random variables, can directly accommodate model uncertainty. The challenge is that the Bayesian posterior distribution is supported on a union of spaces of different dimensions, which makes computation difficult. We review a reversible jump Markov chain Monte Carlo method for sampling from the posterior, and apply this method to provide a Bayesian analysis of simulated and real data.

Acknowledgements: This work is supported by the UIC Liberal Arts and Sciences Undergraduate Research Initiative (LASURI) program. The author is also grateful for the guidance provided by the project's faculty advisor, Professor Ryan Martin.

1 Introduction

Model selection is an inherently difficult problem that is relevant to all fields involving quantitative analysis of data. Intuitively, the problem is that a simple model is preferable to a complex one, but the complex model necessarily fits the data better. Common approaches to this problem augment the model fit criterion with a penalty against those models having more parameters, thus balancing model fit with model complexity. But this approach does not provide any meaningful data-dependent measure of the uncertainty in the model selection step. Uncertainty in the model selection step is an analogous concept to uncertainty in the value of a parameter. Rather than estimate which model is the correct model and operate under the assumption that our estimate is the truth, we acknowledge the possibility that our estimate is incorrect, and instead only assign it some degree of belief. A Bayesian approach, on the other hand, can handle this model uncertainty by treating the model itself as a parameter with an associated prior distribution and by updating this prior distribution to a posterior distribution via Bayes's formula. A challenge in this Bayesian framework is posterior computation, and this is further exaggerated in the case where the model and, therefore, the dimension of the parameter space, varies. For this we employ a reversible jump Markov chain Monte Carlo method [5, 8, 10], which is presented at a level that should be understandable by one with an undergraduate level background in statistics, to sample from the posterior distribution of the model and the model-specific parameters. Output of this procedure includes a probabilistic summary of the information in the data concerning the model itself.

In this paper, the focus is on a particular class of time series models, called autoregressive models. These models depend on an "order" which describes the memory of the process, that is, the number of previous observations that next observation depends on. We introduce these autoregressive models in Section 2. After a brief review of the general Bayesian approach in Section 3, we proceed to describe a Bayesian analysis of data coming from an autoregressive model with unknown order and the associated challenges. Section 4 reviews a reversible jump Markov chain Monte Carlo method to simulate from the complicated posterior distribution, and Section 5 shows applications of this method to simulated and real data examples. Section 6 outlines some ideas for future research along these lines.

2 Time series and the autoregressive model

We begin by considering data that takes the form of an ordered sequence of observations $\{Y_1, Y_2, ..., Y_n\}$ over a period of time. When confronted with data of this nature, one often finds that the usual assumption of independence between observations is violated. Models that handle this dependence are referred to as time series models.

Of particular interest is a specific type of time series that has several statistical features that are not affected by time. A process Y_t is said to be weakly stationarity if its mean and variance do not depend on t and the correlation between Y_t and Y_s depends only on time lag $|t - s|$. A simple model that satisfies the weakly stationary condition is an autoregressive model. The AR model regresses the distance between Y_t and the mean of the process on the p previous terms' $Y_{t-1}, \ldots, Y_{t-p}$ deviation from the mean. By assuming Gaussian error terms (of course other distributions are possible), once the order has been specified, the model parameters can be estimated by maximizing the likelihood function shown below [3]

$$\left(\frac{1}{2\pi \det \Sigma}\right)^{n/2} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}$$

where Σ is the covariance matrix implied by the autoregressive model's dependence structure

What is problematic in this situation is that when comparing likelihoods across nested models of different sizes, models using additional parameters will almost always appear preferable. A typical way of handling this is to introduce a penalty on either the number of parameters or the magnitude of the parameters (where the latter can potentially perform model selection if parameters being set to 0 is interpreted as not using them). In the context of time series models, perhaps the most popular form of regularization is what is known as Akaike Information Criterion, in which an estimate of the Kullback-Leibler divergence between the actual data generating process (which is unknown) and the predictive density is minimized [2]. While this approach is reasonable if one were only interested in selecting a single "best" model from the set of competing models, it is unable to provide answers to additional questions of interest beyond selecting the best model. For example, suppose one was interested in expressing *how* certain they were in a particular model being correct compared to others.

3 Bayesian analysis

The distinguishing feature of the Bayesian framework is that inference is based strictly on the rules of ordinary probability; see, for example, the text by Gosh, Delampady, and Samanta [7]. This is accomplished by introducing a prior distribution for unknown parameter vector θ, including possibly the model index p; see the text by Clyde and George [6]. This prior distribution encodes one's prior beliefs about the parameters before data is observed. Then the data is used to update the prior to a posterior, the conditional distribution of the parameter vector given the data, via Bayes's formula. That is, if θ is the unknown parameter, assigned a prior distribution with density $\pi(\theta)$, then the posterior distribution for θ, given data Y, has density $\pi_Y(\theta)$ given by

$$\pi_Y(\theta) = \frac{L(\theta)\pi(\theta)}{\int L(\theta)\pi(\theta)\, d\theta},$$

where $L(\theta)$ is the likelihood function, which is the density function of Y, given θ, but treated as a function of θ. Uncertainty about the parameter θ is encoded in the posterior distribution, and inference is carried out via certain summaries of the posterior, for example, the posterior mean acts like a point estimator, similar to the maximum likelihood estimator.

In order to address the unknown number of parameters in our autoregressive problem, we consider a hierarchical model, in which the model itself is a parameter. In other words, p is treated as a random variable, where $p = 1$ corresponds to an AR(1) model, $p = 2$ corresponds to an AR(2) model, and so forth. A prior $\pi(p)$ is specified over the collection of models, and within each model a prior $\pi(\theta_p|p)$ is specified, where θ_p are the parameters specific to the AR model of order p. We note that the additional prior distributions within models of greater size has an effect somewhat similar to AIC, in that models with additional parameters must earn their densities with large increases in likelihood. The key difference though, is that under this approach, the posterior will yield genuine probabilities. The "best" model can still be selected using some feature of the distribution, for example, the mode, however we also now have a distribution that can be used to express our belief in a certain model or collection of models being the true model.

Unfortunately the actual mechanics of performing the analysis proposed above end up being rather complex. Even after choosing relatively simple priors, we find that the expression for posterior distribution is sufficiently complex such that its proportionality constant cannot be evaluated analytically, and neither can any posterior summary of interest. So, we will need to rely on numerical methods—in particular, Monte Carlo—to summarize the posterior. To make matters worse, this distribution varies in dimension over different values of p. That is, the posterior distribution is supported on a set consisting of a union of spaces of varying dimension.

4 Reversible Jump MCMC

In order to obtain a random sample from a distribution given only its density up to a constant in a computationally feasible manner, we direct our attention towards Markov chain Monte Carlo. In this approach, we will simulate a Markov chain that has a stationary distribution identical to the distribution we wish to sample from. While there are some downsides to this, namely that some dependence has been introduced between points in the sample, we will find that the ease in computation that this approach provides is enough to justify this disadvantage.

When the dimension of the distribution to be sampled from is fixed, the Metropolis–Hastings method provides a simple way to construct a Markov chain with the desired stationary distribution [4]. This is accomplished by splitting the transition kernel of the chain, $P(\theta \to \theta')$ into two steps, the proposal step $g(\theta \to \theta')$, and the acceptance step $A(\theta \to \theta')$. In the proposal step, θ' is drawn from specified distribution $g(\theta \to \theta')$. As a simple example, one could choose $\theta' \sim \mathcal{N}(\theta, I\sigma^2)$. In the acceptance step, we will either accept or reject proposed θ' according to a specified distribution $A(\theta \to \theta')$. By accept it is meant that the proposed state θ' is taken to be the next state of the chain, and by reject it is meant that the current state θ is taken to be the next state of the chain. By choosing g and A such that detailed balance condition $\pi(\theta)P(\theta \to \theta') = \pi(\theta')P(\theta' \to \theta)$ for all $\theta, \theta' \in \Theta$, is satisfied, elementary results of probability theory can be used to show that the chain has distribution π as its limiting distribution.

The challenge is to extend the procedure presented above in a way such that the chain is able to expand and contract in dimension in order to be able to sample from our posterior which varies in dimension. This will be accomplished through an extension to Metropolis—Hastings known as reversible jump Markov chain Monte Carlo. While the fully generalized version of the algorithm is somewhat complicated, our problem's structure has certain features that can be exploited to simplify the procedure, namely that the set of competing models has a natural ordering, the value of p in each model. A simplified version of the algorithm presented in a 2003 paper by Brooks [5] is presented, and then how it can be set up to be applied to our specific problem.

The chain's movement will consist of two distinct forms, one being how it is to move within a model, which can be accomplished through usual fixed dimension MCMC methods, and the other being how it is to move across models, which is the new part. In order to accomplish moving across models, the proposal step is split into two parts, the first being some scheme to decide which model to consider moving to, and the second being some mechanism that allows parameters to be filled in to match the dimension of proposed models of greater size, or, allows parameters to be deleted to match the dimension of proposed models of lesser size. Both of these steps must be done with great caution, in order to ensure that the detailed balance condition remains satisfied. This will be achieved by ensuring that given a proposed jump from one model to another, we can always recover what would have caused the proposed model to propose moving to the current model, hence the name reversible jump.

The condition rendered above can be accomplished by carefully describing how the chain is to move up in dimension, and then whenever it proposes to move down, considering the reverse move, which is up. Moves up in dimension are performed as follows. Suppose the chain is currently at state $\theta = (\theta_1, ..., \theta_p)$, with corresponding model M. It then selects a model M' for the proposal θ' according to some distribution $J(\theta \to \theta')$. The parameters θ' are then filled in by drawing a vector v of size $p' - p$ from some specified distribution D, appending v to theta to form (θ, v), and then sending (θ, v) through diffeomorphism $h : \mathbb{R}^{p'} \to \mathbb{R}^{p'}$ to finally obtain $\theta' = h(\theta, v)$. [9]

The proposed θ' is then accepted with probability

$$A(\theta \to \theta')_{p<p'} = \min\left\{1, \frac{\pi(\theta')}{\pi(\theta)} \frac{J(\theta' \to \theta)}{J(\theta \to \theta')} \frac{1}{f_D(v)} \Big| \det\big(\frac{\partial h(\theta, v)}{\partial(\theta, v)}\big)\Big|\right\}$$

where f_D is the density function of distribution D.

In the case where $p > p'$, the proposed move is accepted with reciprocal probability of the acceptance probability of the reverse move. In other words, the dimensions are matched by simply deleting parameters, and instead of drawing v, figuring out what draw of v would have produced the current state given the proposed state. The proposal is then accepted with probability

$$A(\theta \to \theta')_{p>p'} = \min\left\{1, \left\{\frac{\pi(\theta)}{\pi(\theta')} \frac{J(\theta \to \theta')}{J(\theta' \to \theta)} \frac{1}{f_D(v)} \Big| \det\big(\frac{\partial h(\theta, v)}{\partial(\theta, v)}\big)\Big|\right\}^{-1}\right\}$$

where $v = (h^{-1}(\theta')_{p'+1}, ..., h^{-1}(\theta')_p)$

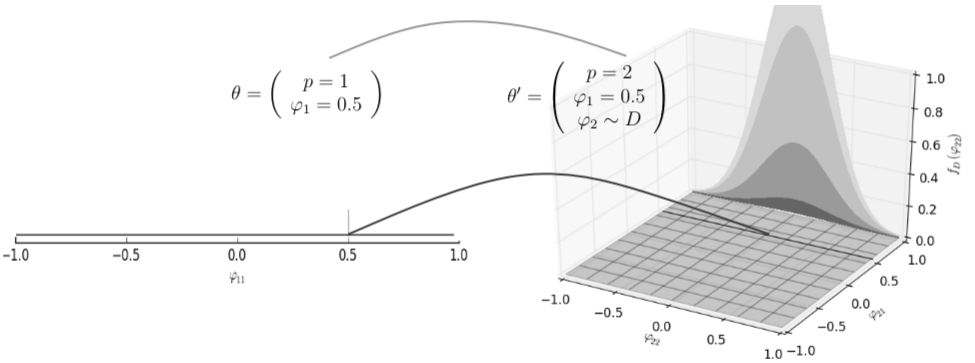
$\theta = \begin{pmatrix} p=1 \\ \varphi_1 = 0.5 \end{pmatrix}$
$\theta' = \begin{pmatrix} p=2 \\ \varphi_1 = 0.5 \\ \varphi_2 \sim D \end{pmatrix}$
−1.0
−0.5
0.0
0.5
1.0
φ_{11}
1.0
0.8
0.6
0.4
0.2
0.0
$f_D(\varphi_{22})$
φ_{22}
φ_{21}

Figure 1: Illustration of chain moving from state of dimension 1 to state of dimension 2

Fortunately in our autoregressive problem we will be able to choose rather simple specifications for $J(\theta \to \theta')$, D, and h. Intuitively, the distribution of p should be approximately unimodal and adjacent models should have relatively similar parameters. We thus are able to "get away" with the following simple proposal scheme. Given that we are at state $\theta = \{p, \sigma_\epsilon, \varphi_1, ..., \varphi_p\}$, we propose moving up in dimension by 1 with probability $\frac{1}{2}$, else moving down in dimension by 1. Moves that are out of bounds are rejected with probability 1. For up moves we draw $v \sim \mathcal{N}(0,1)$ and let $h(\theta, v) = (\theta, v)$. In other words, v is appended to the current set of parameters to form proposal $\theta' = \{p+1, \sigma_\epsilon, \varphi_1, ..., \varphi_p, v\}$. Again, we are able to do this due to the similarity in parameters of close models, if this were not the case then we would greatly benefit from carefully choosing a different h, see the first example in Greens' 1995 paper [8] for an example of a non-identity h. By inserting these choices into the above acceptance probability, θ' is accepted with probability

$$A(\theta \to \theta')_{p \to p+1} = \min \left\{ 1, \frac{\pi(\theta')}{\pi(\theta)} \frac{1}{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{v^2}{2}\right)} \right\}$$

For down moves, φ_p is removed from the current set of autoregressive parameters to form proposal $\theta' = \{p-1, \sigma_\epsilon, \varphi_1, ..., \varphi_{p-1}\}$, which is accepted with probability

$$A(\theta \to \theta')_{p \to p-1} = \min \left\{ 1, \frac{\pi(\theta')}{\pi(\theta)} \frac{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\varphi_p^2}{2}\right)}{1} \right\}$$

After each jump step, time is spent exploring the model-specific parameter space with some vanilla Markov chain Monte Carlo scheme. In our implementation we choose symmetric random walk Metropolis–Hastings. Computer code to carry out the computations discussed in this section is available at https://bitbucket.org/nroger4/bayesian-estimation-in-autoregressive-models-using-reversible/src.

Table 1: Maximum likelihood fit

p	$\hat{\varphi}_1$	$\hat{\varphi}_2$	$\hat{\varphi}_3$	$\hat{L}$	AIC
1	0.2535	—	—	-247.53	501.05
2	0.2909	-0.1486	—	-245.44	498.87
3	0.2957	-0.1575	0.0310	-245.35	500.69

5 Examples

The sampler is applied to three datasets. The first two are simulated AR processes of length $n = 1000$, where both the model and the model-specific parameters are known. It is important to test on data of this nature in order to verify correctness of the implementation of the sampler. Figures 2–3 present the results of running the sampler on two different simulated data sets. In both cases, the trace plots show that the Markov chains are mixing well and that posterior distribution for the model-specific parameters are concentrating around their true values. Perhaps more importantly, the posterior distribution for the model size is clearly identifying the actual model size.

The third example is based on real data, in particular, a differenced series of monthly Treasury bill interest rates from 1950 to 2000; see Section 9.16.1 in the text by Ruppert [13]. Table 1 shows the results of a model fit based on maximum likelihood and AIC. The AIC model selection criterion identifies the AR(2) model as best, but the AR(1) and AR(3) models have similar and only slightly larger AIC scores; note that in the $p = 3$ model, $\hat{\varphi}_3$ is close to zero, suggesting that the AR(3) model is at least slightly too big. Figure 4 shows the results of the posterior simulation for these data. Of particular interest is that the posterior distribution for p is concentrated mostly on $p = 2$ and $p = 1$, assigning little probability to the AR(3) model. This allows for a more complete description of our uncertainty about the model compared to the AIC selection rule.

6 Directions of Future Research

When choosing the user specified components of the reversible jump sampler for our autoregressive problem, the structure of the family of models allowed us to successfully sample from the posterior with a simple jump proposal mechanism. In the case where we are comparing models that do not exhibit such structure, choosing $h(\theta, v)$ will be a difficult task. In fixed-dimension MCMC, there are theoretical results regarding ideal acceptance rate of a chain as a function of dimension. This allows the sampler to modify its proposal distribution as it runs in order to better match the ideal acceptance rate for the given dimension. This must be done with great caution, as it violates the Markov property, in that future states of the chain now have dependence on all previous states. Perhaps this methodology can be adapted to the reversible jump sampler in some manner.

Along a different line, there have been some recent developments on alternatives to clas-

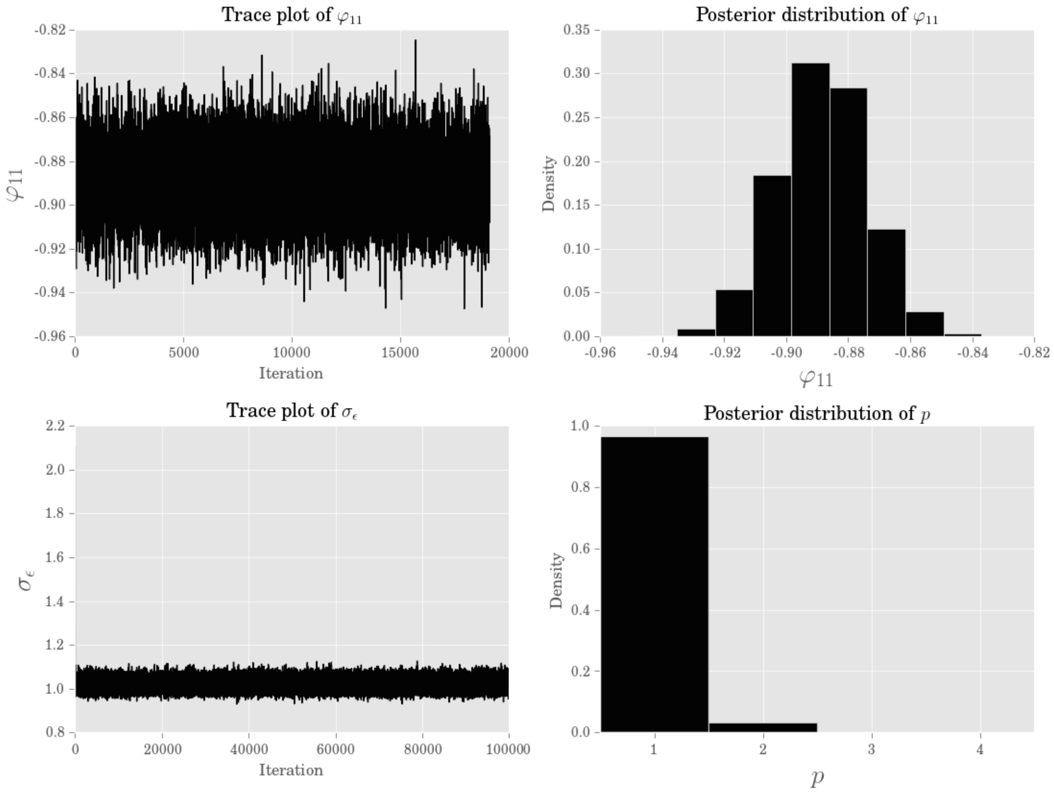

Figure 2: Simulated AR(1) process with $\varphi = (-0.9)$ $\sigma_\epsilon = 1.0$

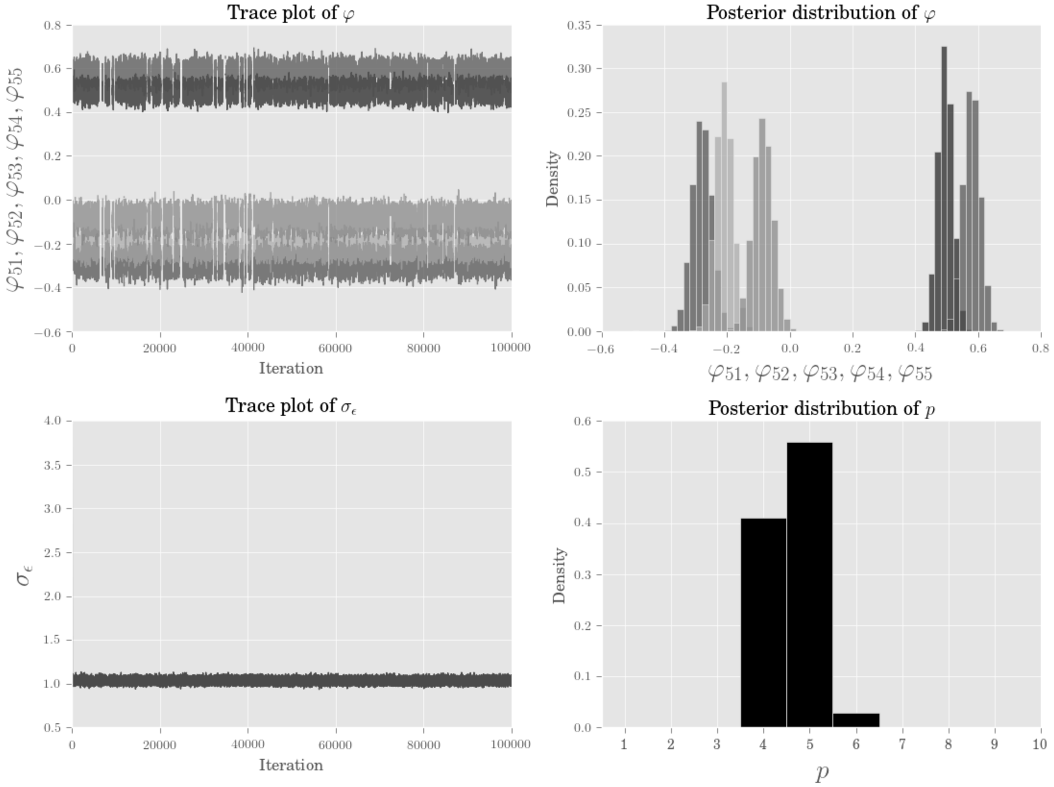

Figure 3: Simulated AR(5) process with $\varphi = (-0.3, -0.2, 0.5, 0.6, -0.1)$, $\sigma_\epsilon = 1.0$

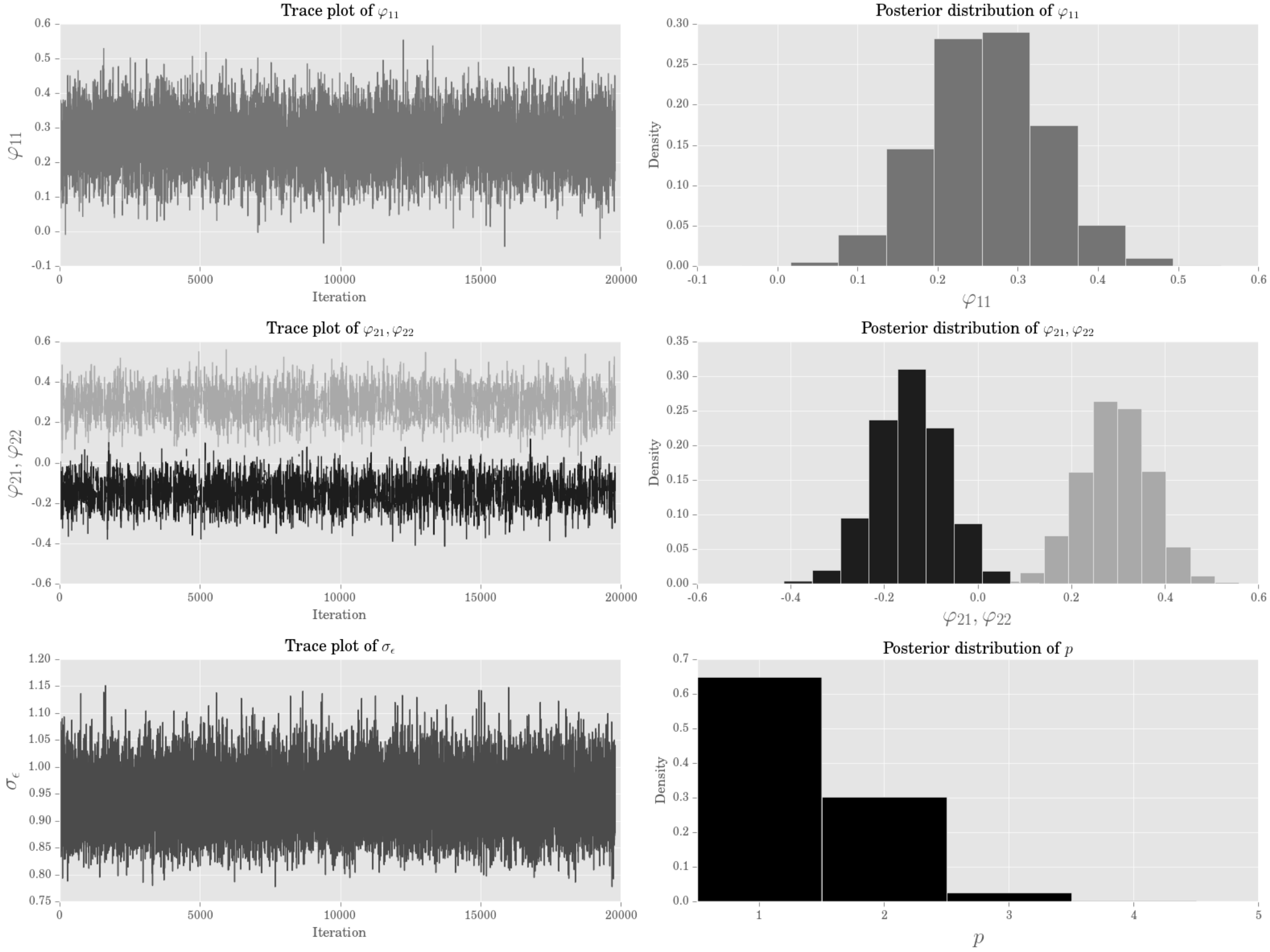

Figure 4: Posterior simulation output from the real-data example.

sical and Bayesian statistical inference. In particular, the *inferential model* (IM) approach in [11] provides valid probabilistic inference without requiring specification of a prior distribution. These ideas were applied recently in [12] to the variable selection problem, with good results. We plan to explore what IMs have to offer compared to the complicated Bayesian approach discussed here.

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium of Information Theory*, 267–281.

[2] Akaike, H. (1974). A new look at the statistical model identification, IEEE Transactions on Automatic Control 19 (6), 716723.

[3] Brockwell, P.J. & Davis, R.A. (1996). Introduction to Time Series and Forecasting, Springer-Verlag, New York .

[4] Brooks, S., Gelman, A., Jones, G., and Meng, X. (2010). *Handbook of Markov Chain Monte Carlo*, Chapman and Hall/CRC Press.

[5] Brooks, S., Giudici, P. and Roberts, G. (2003). Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions. *J. R. Statist. Soc. B*, **65**, 3–55.

[6] Clyde, M. and George, E. (2004). Model uncertainty. *Statist. Sci.*, **19**, 81–94.

[7] Ghosh, J. K., Delampady, M., and Samanta, T. (2006). *An Introduction to Bayesian Analysis*, Springer.

[8] Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

[9] Green, P., and Hastie, D. (2009). Reversible jump MCMC. Available at: http://www.maths.bris.ac.uk/ peter/papers/rjmcmc_20090613.pdf. Accessed: November 28 2015.

[10] Lange, K. (2010). *Numerical Analysis for Statisticians*. Springer, New York, 2nd edition.

[11] Martin, R. and Liu, C. (2013). Inferential models: a framework for prior-free posterior probabilistic inference. *J. Amer. Statist. Assoc.*, **108**, 301–313.

[12] Martin, R., Xu, H., Zhang, Z., and Liu, C. (2014). Optimal prior-free probabilistic variable selection in Gaussian linear regression. Submitted manuscript, `arXiv:1412.5139`.

[13] Ruppert, D. (2011). *Statistics and Data Analysis for Financial Engineering*. Springer, New York.