BAYESIAN MODEL SELECTION FOR TIME SERIES USING MARKOV CHAIN MONTE CARLO

Paul T. Troughton and Simon J. Godsill

Signal Processing and Communications Group, Department of Engineering,
University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, England
ptt10@cam.ac.uk   sjg@eng.cam.ac.uk

ABSTRACT

We present a stochastic simulation technique for subset selection in time series models, based on the use of indicator variables with the Gibbs sampler within a hierarchical Bayesian framework. As an example, the method is applied to the selection of subset linear AR models, in which only significant lags are included. Joint sampling of the indicators and parameters is found to speed convergence. We discuss the possibility of model mixing where the model is not well determined by the data, and the extension of the approach to include non-linear model terms.

1. INTRODUCTION

Until recently, research into time series modelling has concentrated on those models which are analytically convenient, without necessarily justifying underlying assumptions such as linearity. With rapidly increasing computing power, it is now possible to consider a much wider range of models, including hybrids containing terms from several non-linear model families. The problem becomes one of *subset selection* — we wish to select the best subset of terms from the pool available.

We will take a Bayesian approach, since this leads to consistent model selection criteria and avoids the need to introduce explicit penalisation of complex models. By selecting models on the basis of posterior probabilities, we also have the opportunity to incorporate any prior knowledge.

We wish to fit to a time series $\mathbf{x}$ a model which consists of a number of terms with (possibly vector) parameters $\boldsymbol{\theta}_1, \dots \boldsymbol{\theta}_P$. We associate a binary indicator γ_i with each term such that if $\gamma_j = 1$ then the term with parameter $\boldsymbol{\theta}_j$ is included in the model; otherwise it is excluded. We gather any parameters which are common to all models into $\boldsymbol{\phi}$. The most probable combination of terms, and hence the model we wish to select, as represented by $\boldsymbol{\gamma}$, is then:

$$\operatorname*{argmax}_{\boldsymbol{\gamma}}\Big(p(\gamma_1, \gamma_2, \dots \gamma_P \mid \mathbf{x}) = \int_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots \boldsymbol{\theta}_P, \boldsymbol{\phi}} \cdots \int p(\boldsymbol{\theta}_1, \gamma_1, \boldsymbol{\theta}_2, \gamma_2, \dots, \boldsymbol{\phi} \mid \mathbf{x}) \, d\boldsymbol{\theta}_1 \cdots d\boldsymbol{\theta}_P \, d\boldsymbol{\phi}\Big) \quad (1)$$

If we have a pool of P candidate terms, there are 2^P possible combinations. For sizeable P, it becomes impractical to evaluate the probability of *all* subsets. To avoid this, there is a variety of suboptimal search algorithms (see *e.g.* [1, 2]). It is possible that our posterior distributions will be multimodal. This can cause problems with search algorithms, as they tend to stop at local maxima. Hence we will concentrate on stochastic methods.

In [3], an earlier stochastic simulation method is generalised to produce, for any given distribution π, an ergodic Markov chain which has π as a limiting distribution. The Gibbs sampler [4] is a special case of this Metropolis-Hastings algorithm, in which each variable is sampled in rotation from its fully conditional density:

$$\begin{aligned}
\boldsymbol{\psi}_1^{(n+1)} &\sim p(\boldsymbol{\psi}_1 \mid \boldsymbol{\psi}_2^{(n)}, \boldsymbol{\psi}_3^{(n)}, \dots \boldsymbol{\psi}_d^{(n)}, \mathbf{x}) \\
\boldsymbol{\psi}_2^{(n+1)} &\sim p(\boldsymbol{\psi}_2 \mid \boldsymbol{\psi}_1^{(n+1)}, \boldsymbol{\psi}_3^{(n)}, \dots \boldsymbol{\psi}_d^{(n)}, \mathbf{x}) \\
&\vdots \\
\boldsymbol{\psi}_d^{(n+1)} &\sim p(\boldsymbol{\psi}_1 \mid \boldsymbol{\psi}_1^{(n+1)}, \boldsymbol{\psi}_2^{(n+1)}, \dots \boldsymbol{\psi}_{d-1}^{(n+1)}, \mathbf{x}) \qquad (2)
\end{aligned}$$

If we allow the $\boldsymbol{\psi}_i$ to be multivariate, we have a multi-move Gibbs sampler.

Returning to equation (1), a Markov chain can be constructed which moves around the model space by sampling both the indicators and the other model parameters to produce a sequence of states $\boldsymbol{\gamma}^{(1)}, \boldsymbol{\gamma}^{(2)}, \dots,$ which converges in the limit to produce (dependent) samples from the posterior $p(\boldsymbol{\gamma} \mid \mathbf{x})$, thereby performing numerically the integration of equation (1) [5]. From these sampled states, we can obtain Monte Carlo estimates of the marginal posterior density of the indicators.

Both [6] and [7] use variations on this approach for statistical model selection problems, but do not discuss time series. [8] works with subset AR models using the method of [6], in which disabled terms are never completely excluded, but are instead given a narrow prior. We expand on the approach of [7], as completely removing disabled terms gives computational advantages.

It has been argued [6] that, as we are only interested in the subsets with highest posterior probability, rather than evaluation of the full posterior, a run of length $\ll 2^P$ should suffice; for all but the most degenerate multimodal posteriors, this seems reasonable.

If variables are not independent, the Gibbs sampler tends to converge slowly [9]. Since there is likely to be

Work supported by the Engineering and Physical Sciences Research Council.

strong interdependence between the indicator and parameter(s) of each term, we speed convergence by sampling *jointly* from the indicators and their associated parameters, in a similar manner to that used for impulse detection in [10, 11].

There is also interdependence between the parameters and indicators of different terms. We can address this by multivariate sampling of the indicators, in blocks of size Q. Each iteration then requires the evaluation of the conditional for 2^Q combinations of terms. Varying Q allows a trade-off between the number of iterations required for convergence and the computational complexity of each iteration.

Following from [12], we sample the indicators in random order, but sample the different types of component in a fixed sequence.

2. EXAMPLE

We now illustrate this method with a simple linear model.

2.1. Subset AR model

The subset autoregressive model [13] with maximum order P can be represented in terms of parameters a_i:

$$x_t = e_t + \sum_{i=1}^{P} x_{t-i}\, a_i\, \gamma_i \tag{3}$$

where e_t is an i.i.d. Gaussian excitation sequence with constant variance. With appropriate matrix and vector definitions [14], the conditional likelihood can be expressed as:

$$p(\mathbf{x}_1 \mid \mathbf{a}, \boldsymbol{\gamma}, \sigma_e, \mathbf{x}_0) = (2\pi\sigma_e^2)^{-\frac{N-P}{2}} \exp\left(-\tfrac{1}{2}\sigma_e^{-2}\, \|\mathbf{x}_1 - \mathbf{X}(\mathbf{a}\cdot\boldsymbol{\gamma})\|^2\right) \tag{4}$$

where $\mathbf{x}_0$ contains the first P elements of $\mathbf{x}$, and $\mathbf{x}_1$ the remainder.

2.2. Priors

We use a Bernoulli prior for the indicators, $p(\gamma_i = 1) = \alpha$. For the AR parameter values, we use a convenient prior: independent zero-mean univariate Gaussians, all of variance σ_p. Since the noise variance is a scale parameter, we use a Jeffreys' prior. With suitable bounds, this can be made proper. Alternatively, an Inverse Gamma prior could be used on σ_e^2.

2.3. Conditional distributions

We have two types of sampling step:

$$\mathbf{a}_u, \boldsymbol{\gamma}_u \sim p(\mathbf{a}_u, \boldsymbol{\gamma}_u \mid \mathbf{x}, \mathbf{a}_k, \boldsymbol{\gamma}_k, \sigma_e) \tag{5}$$

$$\sigma_e \sim p(\sigma_e \mid \mathbf{x}, \mathbf{a}, \boldsymbol{\gamma}) \tag{6}$$

where the subscripts $(\cdot)_u$ and $(\cdot)_k$ denote partitioning into, respectively, those elements corresponding to terms whose indicators are being sampled, and those which are currently being regarded as fixed.

The joint sampling operation of step (5) can be performed in two steps:

$$\boldsymbol{\gamma}_u \sim p(\boldsymbol{\gamma}_u \mid \mathbf{x}, \mathbf{a}_k, \boldsymbol{\gamma}_k, \sigma_e) \tag{7a}$$

$$\mathbf{a}_u \sim p(\mathbf{a}_u \mid \mathbf{x}, \mathbf{a}_k, \boldsymbol{\gamma}, \sigma_e) \tag{7b}$$

Note that step (7a) is *not* conditional on $\mathbf{a}_u$. We find this first, discrete, distribution from the likelihood by repeated application of Bayes' rule and by marginalising $\mathbf{a}_u$, giving:

$$p(\boldsymbol{\gamma}_u \mid \mathbf{x}, \mathbf{a}_k, \boldsymbol{\gamma}_k, \sigma_e) \propto (2\pi\sigma_e^2)^{-\frac{N-P}{2}} \sqrt{\frac{|\mathbf{C}_s|}{|\mathbf{C}_p|}} \exp\left(-\tfrac{1}{2}(\boldsymbol{\mu}_p^T \mathbf{C}_p^{-1} \boldsymbol{\mu}_p - \boldsymbol{\mu}_s^T \mathbf{C}_s^{-1} \boldsymbol{\mu}_s)\right) \cdot \alpha^{n_1}(1-\alpha)^{(l-n_1)} \tag{8}$$

where l is the dimension of $\boldsymbol{\gamma}_u$, n_1 is the number of components of $\boldsymbol{\gamma}_u$ which are 'on', $\boldsymbol{\mu}_p$ and $\mathbf{C}_p$ are the mean vector and covariance matrix of the Gaussian prior $p(\mathbf{a}_u)$, and:

$$\mathbf{C}_s = (\sigma_e^{-2}\mathbf{X}_u{}^T\mathbf{X}_u + \mathbf{C}_p^{-1})^{-1} \tag{9}$$

$$\boldsymbol{\mu}_s = \mathbf{C}_s(\sigma_e^{-2}\mathbf{X}_u{}^T(\mathbf{x} - \mathbf{X}_k\mathbf{a}_k) + \mathbf{C}_p^{-1}\boldsymbol{\mu}_p) \tag{10}$$

In these terms, the distribution required for step (7b) is simply the multivariate Gaussian

$$p(\mathbf{a}_u \mid \mathbf{x}, \mathbf{a}_k, \boldsymbol{\gamma}, \sigma_e) \propto \mathrm{N}(\mathbf{a}_u \mid \boldsymbol{\mu}_s, \mathbf{C}_s) \tag{11}$$

Sampling all of $\mathbf{a}$ is a simple operation, based on equation (11) with $\mathbf{a}_u = \mathbf{a}$ and $\mathbf{a}_k$ empty. Occasionally including this step:

$$\mathbf{a} \sim p(\mathbf{a} \mid \mathbf{x}, \boldsymbol{\gamma}, \sigma_e) \tag{12}$$

can further reduce the effect of interdependence between AR parameters.

Finally, the fully conditional distribution of the noise variance (eq. 6) is found to be an Inverse Gamma distribution, for which well-known sampling methods exist.

3. RESULTS

The above sampler was implemented, and experiments were performed using both synthetic and real data.

3.1. Synthetic data

800 samples of synthetic data were generated from a subset AR model containing terms of order $\{1, 2, 3, 5, 7\}$. The sampler was run with candidate terms up to order 9. The initial values of the indicators, parameters and noise variance were zero. Indicators were sampled in triples.

Figure 1 shows the results of a typical run of 150 iterations, of which the first 50 were discarded as burn-in. The top plot shows the mean value of each of the indicators, which can be interpreted as the marginal posterior probability of inclusion of each of the model terms. It can be seen that the correct terms come out with clearly higher probability.

An alternative method for choosing a model from this data is to find the *combination* of terms which appears most frequently [6]. This frequency should be an estimate of the subset's posterior probability. The middle plot

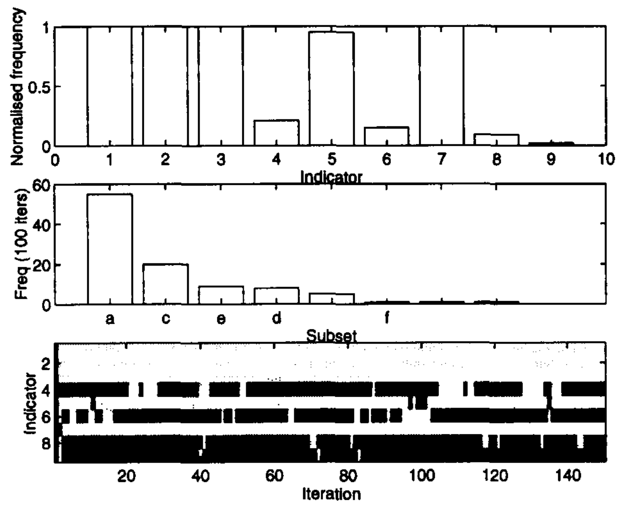

Figure 1: Synthetic data — simulation results after 150 iterations (including 50 iterations burn-in): *(top)* Normalised indicator frequencies; *(middle)* Indicator subset frequencies; *(bottom)* Raw indicator values.

shows that there is a clear favourite; this is the correct model. The labelling corresponds to figure 2.

The bottom plot shows the values of the indicators in each iteration — those indicators which are switched on are shown as white pixels.

It was found that, as is normal in Bayesian inference, the above results were insensitive to variations of σ_p and α over a wide range when the model was well determined by the data.

3.2. Analytic results

To verify that the sequence of states being produced is correct for subset selection, the posterior model probability was evaluated analytically for each of the possible subsets. This exhaustive calculation is feasible only for small P, and requires knowledge of the correct value of σ_e. The same model was used as for §3.1, but this time only 400 samples were generated.

Figure 2 shows histograms generated from the calculated evidence, together with the simulation results (with fixed σ_e) for both a large and a small number of iterations. It can be seen that the long run agrees closely with the evidence, and the short run, although more coarse, would lead to the selection of the same model.

3.3. Audio data

Figure 3 shows 1000 samples from an orchestral recording, together with the values of the AIC and MDL criteria for different orders of non-subset AR models. The AIC would lead to a choice of an AR(20) model, whereas the MDL favours an AR(12) model. The AIC is known to tend to overfit.

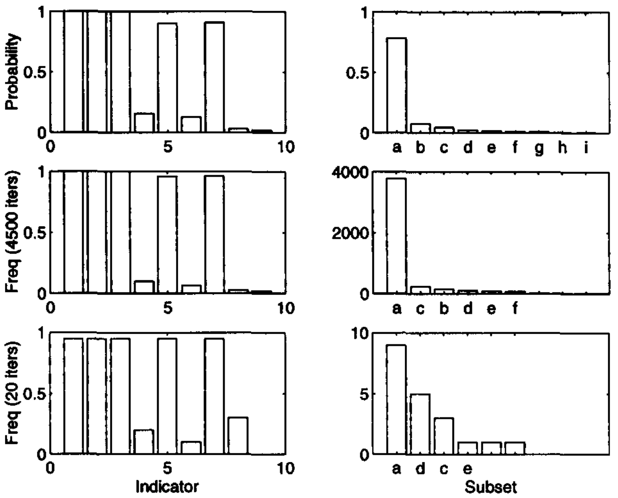

Figure 2: Comparison of *(top)* analytically calculated posterior model probabilities with *(middle)* simulation results from 5000 iterations (including 500 iterations burn-in) and *(bottom)* from 30 iterations (including 10 iterations burn-in).

The figure also gives the results of running the sampler with candidate terms to order 40 for 300 iterations, discarding the first 50 as burn-in. The highest frequency subset, a plain AR(12) model, accounts for some 85% of the post burn-in iterations.

4. NON-LINEAR MODELS AND APPLICATIONS

It is straightforward to extend these methods to include polynomial terms, forming a truncated Volterra series [15]. In the second-order case (with P lags), the modelling equations take the form:

$$x_t = e_t + \sum_{i=1}^{P} x_{t-i}\, a_i\, \gamma_i + \sum_{i=1}^{P}\sum_{j=1}^{P} b_{ij}\, \xi_{ij}\, x_{t-i}\, x_{t-j} \qquad (13)$$

which can be converted into matrix-vector notation as:

$$\mathbf{e} = \mathbf{x}_1 - \mathbf{X}'(\mathbf{a}' \cdot \boldsymbol{\gamma}') \qquad (14)$$

where $\mathbf{a}'$ contains both $\{a_i\}$ and $\{b_{ij}\}$, γ' contains both $\{\gamma_i\}$ and $\{\xi_{ij}\}$, and $\mathbf{X}'$ is a block diagonal matrix. Note that this system is still *linear in the parameters*, and of the same form as equation (4).

It should also be possible to include other non-linear terms, with more parameters, such as thresholds [15].

Having developed a sampler for model selection, we can incorporate extra steps to use the model to produce any required output, such as forecasts or a reconstruction of missing data.

In terms of equation (1), the required output can be included in ϕ. This approach has the advantage, other than simplicity, that, in the event of model uncertainty,

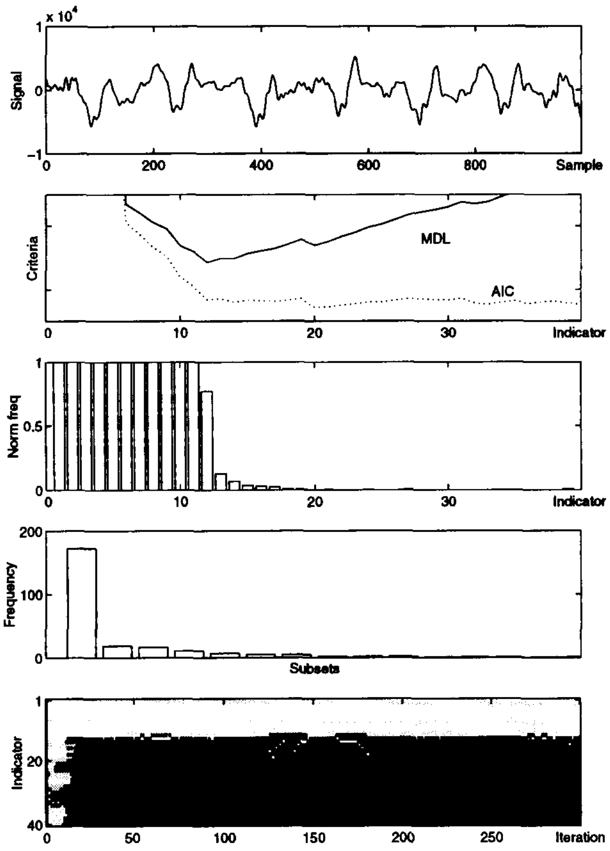

Figure 3: Orchestral recording (from top): *(a)* Signal; *(b)* MDL (solid) and AIC (dotted) values for non-subset AR models; *(c)* Normalised indicator frequencies; *(d)* Indicator subset frequencies; *(e)* Raw indicator values.

the output will be based on processing using *all* the probable models, rather than just the one with highest posterior probability, *i.e.*

$$p(\phi \mid \mathbf{x}) = \int_{\theta_1,\theta_2,\ldots\theta_P} \cdots \int \sum_{\gamma_1,\gamma_2,\ldots\gamma_P} \cdots \sum p(\boldsymbol{\theta}_1,\gamma_1,\ldots \mid \mathbf{x})\, d\boldsymbol{\theta}_1 \cdots d\boldsymbol{\theta}_P \qquad (15)$$

This approach has been used with a linear model for signal reconstruction in the presence of impulsive and continuous noise [16]; the ability to incorporate non-linear model terms should make it possible to reconstruct audio which has suffered distortion by a poor recording chain [17].

5. CONCLUSIONS

We have shown, using the example of a simple linear model, that this method provides a means of avoiding the 2^P combinatorial explosion associated with subset selection. The MCMC framework has the advantage that it can be applied to models which are not analytically tractable. Furthermore, it allows much flexibility in producing output, and copes elegantly with model uncertainty.

6. REFERENCES

[1] K. J. Pope & P. J. W. Rayner. "Non-linear system identification using Bayesian inference". *Proc. IEEE ICASSP*, IV:457–460, 1994.

[2] G. M. Furnival & R. W. Wilson, Jr. "Regressions by leaps and bounds". *Technometrics*, 16(4):499–511, 1974.

[3] W. K. Hastings. "Monte Carlo sampling methods using Markov chains and their applications". *Biometrika*, 57(1):97–109, 1970.

[4] S. Geman & D. Geman. "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images". *IEEE Trans. PAMI*, PAMI-6(6):721–741, 1984.

[5] A. E. Gelfand & A. F. M. Smith. "Sampling-based approaches to calculating marginal densities". *J. Am. Stat. Assoc.*, 85(410):398–409, 1990.

[6] E. I. George & R. E. McCulloch. *Stochastic search variable selection.* In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter, eds., *Markov Chain Monte Carlo in Practice: Interdisciplinary Statistics*, pp. 203–214. Chapman & Hall, 1996.

[7] B. P. Carlin & S. Chib. "Bayesian model choice via Markov chain Monte Carlo". *J. Roy. Stat. Soc. B*, 57:473–484, 1995.

[8] C. W. S. Chen. *On the selection of best subset autoregressive time series models.* Tech. rep., Dept. of Statistics, Feng-Chia Univ., Taiwan, 1996.

[9] L. Tierney. "Markov chains for exploring posterior distributions". *Ann. Stat.*, 22(4):1701–1762, 1994. With discussion.

[10] S. J. Godsill & P. J. W. Rayner. *Robust reconstruction and analysis of autoregressive signals in impulsive noise using the Gibbs sampler.* Tech. rep. CUED/F-INFENG/TR.233, Dept. of Engineering, Univ. of Cambridge, 1995.

[11] C. K. Carter & R. Kohn. "Markov-chain monte-carlo in conditionally gaussian state-space models". *Biometrika*, 83(3):589–601, 1996.

[12] J. A. Stark. *Variable Selection in Data and Signal Modelling.* PhD thesis, Univ. of Cambridge, 1995.

[13] J. T. McClave. "Estimating the order of autoregressive models; the max χ^2 method". *J. Am. Stat. Assoc.*, 73:122–128, 1978.

[14] G. E. P. Box & G. M. Jenkins. *Time Series Analysis: Forecasting and Control.* Holden-Day Series in Time Series Analysis. Holden-Day, revised edition, 1976.

[15] H. Tong. *Non-linear Time Series: A Dynamical System Approach.* Oxford Statistical Science Series. Oxford University Press, 1990.

[16] S. J. Godsill & P. J. W. Rayner. "Robust noise reduction for speech and audio signals". *Proc. IEEE ICASSP*, 1996.

[17] K. J. Mercer. *Identification of Distortion Models.* PhD thesis, Univ. of Cambridge, 1993.