VARIABLE SELECTION BY A REVERSIBLE JUMP MCMC APPROACH

Petar M. Djurić *
Department of Electrical Engineering
State University of New York at Stony Brook
Stony Brook, NY 11794, USA
Tel: +1 516 632-8423; fax: +1 516 632-8494
e-mail: djuric@sbee.sunysb.edu

ABSTRACT

In this paper we address the problem of selecting the best subset of predictors in linear models from a given set of predictors. In computing the posterior probabilities of the various models, we propose to use the method of reversible jump Markov chain Monte Carlo sampling which cyclicly sweeps through the set of possible predictors and includes or removes them from the model one at a time. Special emphasis is given to a scheme that does not require sampling of the model coefficients and is based on predictive densities. Numerical results are provided that show the performance of the proposed approach.

1 INTRODUCTION

The problem of variable selection is an old one, and it typically arises during model building. The most widely used models in practice are linear regressions, where a set of observed data is described by a family of explanatory variables or predictors with linear relationships. In general then, all the variables are considered as candidates for the data model, and the objective is to choose those that are, indeed, explanatory. In signal processing there are many problems that can be cast as ones of variable selection, including detection of number of reflections in radar or sonar [8], determining the relevant parameters of time varying models such as time-varying autoregressions [7], [12], or choosing the terms of expansion series used in representing nonlinear systems [9], [11].

The literature of variable selection is abundant. Many procedures have been proposed and thoroughly investigated. Some of them have been developed within the classical hypothesis testing framework, others are based on residual sum of square rules, and a third group exploits information criteria (such as the AIC and BIC), [10]. Recently, a new approach has been proposed which uses a Markov chain Monte Carlo sampling (MCMC) technique (in particular Gibbs sampling) to compute the posterior probabilities of the most promising models [4]. The idea there is to put a hierarchical mixture prior on the model coefficients which allows the formation of a posterior used by the Gibbs sampler. From the drawn samples of this distribution the most promising models are readily identified as the ones which appear most frequently. A related Bayesian variable selection method was reported in [2].

Here we propose a different approach, based on a reversible jump MCMC sampling scheme [6]. Each move of the sampler represents a removal from or inclusion of a predictor to the model. The sampler randomly sweeps the set of predictors and at the end of the sweep, the visited model is recorded. The most frequently visited model is the 'best' model. To reduce the dimensionality of the sampling spaces of the samplers, most or all of the signal and noise parameters are integrated out, and predictive densities are formed. The samplers have been tested, and we provide some numerical results.

2 PROBLEM STATEMENT

Let $\mathcal{H} = \{\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_q\}$, represent a set of predictors whose elements are used to model a data record $\mathbf{y}$ according to

$$\mathbf{y} = \mathbf{h}_1^* \theta_1^* + \mathbf{h}_2^* \theta_2^* + \cdots + \mathbf{h}_p^* \theta_p^* + \boldsymbol{\epsilon} \qquad (1)$$

where $\mathbf{h}_i^* \in \mathcal{H}$, $p < q$, the θ_i^*'s are unknown model coefficients, and $\boldsymbol{\epsilon}$ is an error vector. All the vectors are of dimension $N \times 1$. The main task is to choose the best subset of predictors from the available set, where the size of the subset p is also unknown. Obviously, this is a special type of a model selection problem, where the total number of models is 2^q.

3 VARIABLE SELECTION BY A REVERSIBLE JUMP MCMC APPROACH

In situations where the number of variables is even moderately high, comparison of all the models is computationally prohibitive. In this paper we propose an approach that exploits the reversible jump MCMC sampler, which is a generalization of the standard MCMC method [5]. Its main feature is that it allows for jumps between model spaces in addition to the regular moves within a specific model space. It is based on the Metropolis-Hastings method and is implemented in

*This work was supported by the National Science Foundation under Award No. MIP-9506743.

the usual way, where first, the probability of a move acceptance is computed, followed by a decision whether to accept the move or not. For example, if the current model is $\mathcal{M}_j$ with parameter space $\boldsymbol{\Theta}_j$, and a move m is proposed from $\boldsymbol{\theta}_j \in \boldsymbol{\Theta}_j$ to $\boldsymbol{\theta}_k \in \boldsymbol{\Theta}_k$, where $\boldsymbol{\Theta}_k$ is the parameter space of a higher dimensional model $\mathcal{M}_k$, the acceptance probability is obtained from

$$P_m = \min\left(1, \frac{f(\boldsymbol{\theta}_k|\mathbf{y},\mathcal{M}_k)\; r_m(\boldsymbol{\theta}_k)}{f(\boldsymbol{\theta}_j|\mathbf{y},\mathcal{M}_j)\; r_m(\boldsymbol{\theta}_j)\; g(\mathbf{u})}\left|\frac{\partial\boldsymbol{\theta}_k}{\partial(\boldsymbol{\theta}_j,\mathbf{u})}\right|\right) \tag{2}$$

where $r_m(\boldsymbol{\theta}_j)$ is the probability of choosing move type m when in state $\boldsymbol{\theta}_j$, $\mathbf{u}$ is a random vector independent of $\boldsymbol{\theta}_j$, and $g(\mathbf{u})$ is a density function of $\mathbf{u}$. The proposed point in $\boldsymbol{\Theta}_k$ is obtained by using an invertible function $\boldsymbol{\theta}_k(\boldsymbol{\theta}_j,\mathbf{u})$, and the factor $|\partial\boldsymbol{\theta}_k/\partial(\boldsymbol{\theta}_j,\mathbf{u})|$ is the Jacobian arising from the proposed move. If there is a reverse move, the evaluation of P_m is done similarly. It should be noted that in (2), dimension matching is imposed and detailed balance is retained. It is also important that the so constructed Markov chains are irreducible and aperiodic.

We propose to use the reversible jump sampler for variable selection with a scheme that is implemented in one of two ways. With the first one, no sampling from the parameters spaces takes place, whereas with the second, only a sampling of a visited variable is carried out. Each variable is visited once during a sweep, where a sweep consists of a movement through the complete variable space in a random order. A brief summary of the procedure is given by the following pseudo code:

Beginning of a new sweep
  Randomly choose a new variable
  If variable in the model
    Propose a move for its removal
    If move accepted
      Variable excluded from the model
    Else
      Variable stays in the model
    End
  Else
    Propose a move for its inclusion
    If move accepted
      Variable included in the model
    Else
      Variable stays out of the model
    End
  End
  If all variables have been visited
    Record the model and start a new sweep
  Else
    Go back and randomly choose a new variable
  End
End of sweep

4 IMPLEMENTATION

Here we show how we can implement the proposed scheme without sampling from the parameter space. We use uninformative priors for the unknown parameters, and form predictive densities based on a portion of the data [1], [3]. That is, the data are partitioned into estimation and validation subsets, $\mathbf{y}_e$ and $\mathbf{y}_v$, where the estimation subset is used to obtain proper priors for the parameters, and the validation subset for obtaining the likelihood of the model.

The acceptance probability of the reversible jump MCMC is then computed from

$$P_m = \min\left(1, \frac{f(\mathbf{y}_v|\mathbf{y}_e,\mathcal{M}_k)\; r_m(\mathcal{M}_k)}{f(\mathbf{y}_v|\mathbf{y}_e,\mathcal{M}_j)\; r_m(\mathcal{M}_j)}\right) \tag{3}$$

where the predictive density $f(\mathbf{y}_v|\mathbf{y}_e,\mathcal{M}_k)$ is given by

$$f(\mathbf{y}_v|\mathbf{y}_e,\mathcal{M}_k) = \frac{1}{(2\pi)^{\frac{N-n}{2}}}\frac{\Gamma(\frac{N-p_k}{2})}{\Gamma(\frac{n-p_k}{2})} \times \frac{\left(\frac{\mathbf{y}_e^T\mathbf{P}^{\perp}_{k_e}\mathbf{y}_e}{2}\right)^{\frac{n-p_k}{2}}}{\left(\frac{\mathbf{y}^T\mathbf{P}^{\perp}\mathbf{y}}{2}\right)^{\frac{N-p_k}{2}}}\frac{|\mathbf{H}_{k_e}^T\mathbf{H}_{k_e}|^{\frac{1}{2}}}{|\mathbf{H}_k^T\mathbf{H}_k|^{\frac{1}{2}}} \tag{4}$$

and the predictive density $f(\mathbf{y}_v|\mathbf{y}_e,\mathcal{M}_j)$ is similarly defined. In (4), p_k is the number of predictors in the model $\mathcal{M}_k$, n is the length of the vector $\mathbf{y}_e$, $\mathbf{H}_k$ is an $N \times p_k$ matrix whose columns are the predictors of the model $\mathcal{M}_k$, and $\mathbf{P}_k^{\perp}$ is a projection matrix obtained by

$$\mathbf{P}_k^{\perp} = \mathbf{I} - \mathbf{H}_k\left(\mathbf{H}_k^T\mathbf{H}_k\right)^{-1}\mathbf{H}_k^T. \tag{5}$$

The matrix $\mathbf{H}_{k_e}$ is of dimension $n \times p_k$ and can be defined to be equal to the first n rows of $\mathbf{H}_k$. The projection matrix $\mathbf{P}_{k_e}^{\perp}$ is formed from $\mathbf{H}_{k_e}$ similarly as $\mathbf{P}_k^{\perp}$ from $\mathbf{H}_k$ in (5).

5 SIMULATION RESULTS

The approach was tested on a set of synthesized data generated by models similar to those in [4]. There were three experiments. In the first one the data were obtained from

$$\mathbf{y} = \mathbf{h}_4\theta_4 + \mathbf{h}_5\theta_5 + \boldsymbol{\epsilon} \tag{6}$$

where $\theta_4 = 1$, $\theta_5 = 1.2$, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0},\sigma^2\mathbf{I})$, with $\sigma = 2.5$. The number of candidate predictors q was 5, so the total number of models was 32. The length of the observed data vector was $N = 60$. The predictors were independent and identically distributed according to a multivariate Gaussian density, $\mathbf{h}_i \sim \mathcal{N}(\mathbf{0},\mathbf{I})$, and therefore they were uncorrelated. Thus, the true model included only two variables out of the five. The signal-to-noise ratio for predictor $\mathbf{h}_4$ was almost -8 dB, and for predictor $\mathbf{h}_5$ it was somewhat less than -6 dB. The

model coefficients	probability
θ_4 and θ_5	0.7330
θ_3, θ_4, and θ_5	0.1510
θ_5	0.0360

Table 1: Performance of the reversible jump sampler in experiment one. The entries in the left column represent the nonzero coefficients of the best three models, and the numbers in the right column are the estimated posterior probabilities of the respective models. The nonzero coefficients of the correct model were θ_4 and θ_5.

model coefficients	probability
θ_4 and θ_5	0.4240
θ_3 and θ_4	0.2110
θ_5	0.0890

Table 2: Performance of the reversible jump sampler in experiment two. The entries in the left column represent the nonzero coefficients of the best three models, and the numbers in the right column are the estimated posterior probabilities of the respective models. The nonzero coefficients of the correct model were θ_4 and θ_5.

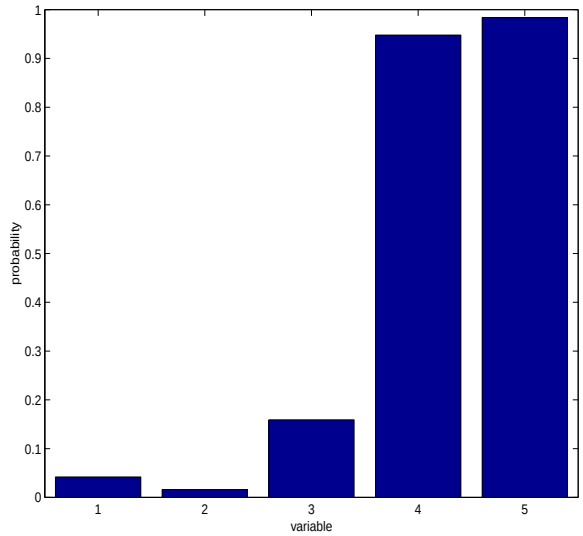

Figure 1: Marginal posterior probabilities of the predictors in experiment one.

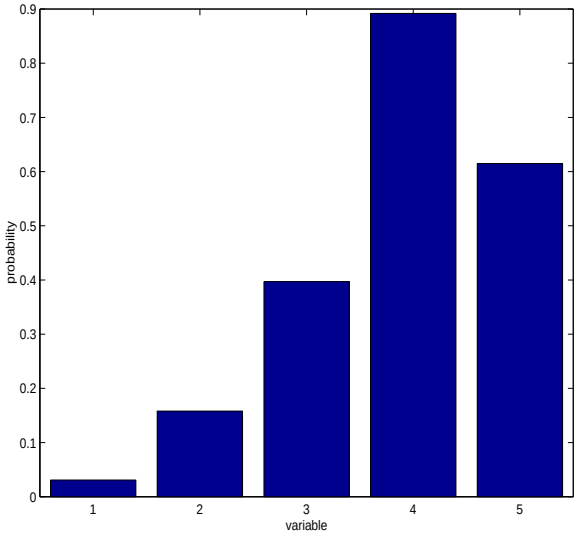

Figure 2: Marginal posterior probabilities of the predictors in experiment two.

least-squares estimate of the model coefficient vector $\boldsymbol{\theta}$ was $\hat{\boldsymbol{\theta}}^T = [-0.0911 \quad -0.2275 \quad 0.5100 \quad 1.0994 \quad 1.3421]$.

We used the reversible jump MCMC sampler where no sampling from the parameter space took place. We ran the sampler for 1020 sweeps, and from the obtained posterior we could construct various estimators. Some results are shown in Table 1 and Figure 1. The entries of the left column in the table are the variables of the most frequently visited models, and the numbers in the right column are the estimated posterior probabilities of these models. The probability of the correct model clearly stands out from the second and third best models. In the figure, the marginal posterior probabilities of the respective predictors are shown. Although these probabilities cannot be used for deciding which predictors to be included in the model, it is striking that the relevant predictors have much higher probabilities than the remaining ones.

In the next experiment the setting was the same except that the third predictor was defined by $\mathbf{h}_3 = \mathbf{h}_5 + 0.15\mathbf{w}$, where $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. This introduced correlation between the third and fifth predictors (with a correlation coefficient equal to 0.989). The least squares estimate of the model coefficient vector $\boldsymbol{\theta}$ was $\hat{\boldsymbol{\theta}}^T = [-0.0805 \quad -0.2064 \quad 0.8032 \quad 1.2116 \quad 0.2300]$. It is interesting to observe that the estimate of θ_3 (whose true value was 0) was greater than the estimate of θ_5 (with true value 1.2). Again, 1020 sweeps were made of which the first 20 were discarded. The results of the reversible jump MCMC sampling are shown in Figure 2 and Table 2.

The strong correlation between $\mathbf{h}_3$ and $\mathbf{h}_5$ notwithstanding, the reversible jump MCMC was able to pick the correct model most frequently. Also, the marginal probability of h_5 is greater than the marginal probability of h_3, regardless of the fact that $\hat{\theta}_5 < \hat{\theta}_3$.

Finally in the third experiment, we simulated a scenario with a fairly large number of predictors. Again the model was

$$\mathbf{y} = \mathbf{H}\boldsymbol{\theta} + \boldsymbol{\epsilon}. \tag{7}$$

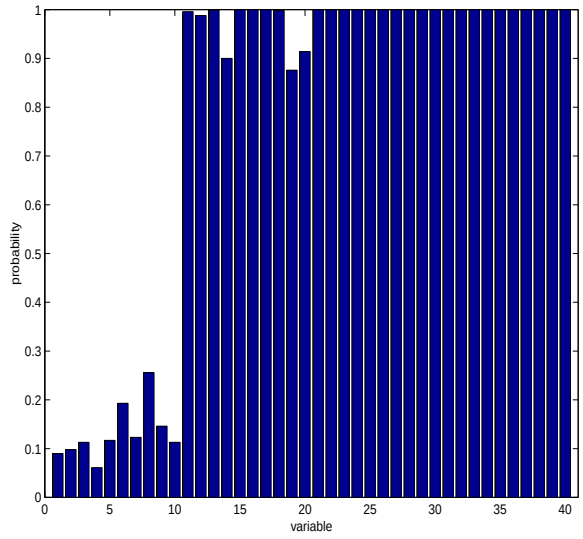

Figure 3: Marginal posterior probabilities of the predictors in experiment three.

We chose $q = 40$, and the number of samples was $N = 120$. The disturbance vector $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$, with $\sigma = 2$. The predictors were defined by $\mathbf{h}_i = \mathbf{u}_i + \mathbf{w}$, where the $\mathbf{u}_i$'s were independent and identically distributed according to a multivariate Gaussian density with mean zero and covariance matrix $\mathbf{I}$. The vector $\mathbf{w}$ was independent from the $\mathbf{u}_i$'s and came from the same distribution. Therefore the pairwise correlation between the predictors was 0.5 The model coefficients took the following values: $\theta_1 = \theta_2 = \cdots = \theta_{10} = 0$, $\theta_{11} = \theta_{12} = \cdots = \theta_{20} = 1$, $\theta_{21} = \theta_{22} = \cdots = \theta_{30} = 2$, and $\theta_{31} = \theta_{32} = \cdots = \theta_{40} = 3$. The number of sweeps was 1020, and the first 20 samples were discarded. The most frequently visitied model was the correct model, and the estimated posterior probability of the correct model was 0.3040. The marginal posterior probabilities of the various predictors in the model are shown in Figure 3. The predictors that are indeed in the model have marginal posterior probabilities very close to one, whereas those which are not in the model, have significantly smaller posterior probabilities.

6 CONCLUSION

Reversible jump MCMC procedures for variable selection in linear models were proposed. The emphasis was on a scheme based on predictive densities which does not require sampling from the parameter space. Results of several tests have been presented, and they have all been excellent. These tests, however, have been carried out on limited number of experiments and with relatively small number of predictors. It is therefore important to investigate the procedure on more difficult tasks, such as those with much larger number of variables. Comparisons with existing methods in terms of accuracy and computational requirements will also be valuable.

References

[1] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*. New York: John Wiley & Sons, 1994.

[2] B. P. Carlin and S. Chib, "Bayesian model choice via Markov chain Monte Carlo methods," Journal of the Royal Statistical Society B, vol. 57, pp. 473-484, 1995.

[3] P. M. Djurić, S. J. Godsill, W. J. Fitzgerald, and P. J. W. Rayner, "Detection and estimation of signals by reversible jump Markov chain Monte Carlo computations," Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Seattle, WA, 1998.

[4] E. I. George and R. McCulloch, "Variable selection via Gibbs sampling," Journal of the American Statistical Association, vol. 88, pp. 881-889, 1993.

[5] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, *Markov Chain Monte Carlo in Practice*. New York: Chapman Hill, 1996.

[6] P. J. Green, "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," Biometrika, vol. 82, pp. 711-732, 1995.

[7] Y. Grenier, "Parametric time-frequency representations," in *Signal Processing*. Eds. J.L. Lacoume, T.S. Durrani, and R. Stora, Elsevier Science Publishers, 1987.

[8] S. M. Kay, *Fundamentals of Statistical Signal Processing, Estimation Theory*, Englewood Cliffs, NJ: Prentice-Hall, 1993.

[9] L. Ljung and T. Söderström, *Theory and Practice of Recursive Identification*. Cambridge, MA: The MIT Press, 1983.

[10] A. J. Miller, *Subset Selection in Regression*. New York: Chapman and Hall, 1990.

[11] J. C. Ralston, A. M. Zoubir, and B. Boashash, "Identification of a class of nonlinear systems under stationary non-Gaussian excitation," IEEE Transaction on Signal Processing, vol. 45, pp. 719-735, 1997.

[12] M. K. Tsatsanis and G. B. Giannakis, "Time-varying system identification and model validation using wavelets," IEEE Transactions on Signal Processing, vol. 41, pp. 3512-3523, 1993.