Proceedings of the 1999 Winter Simulation Conference
P. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, eds.

REGRESSION METAMODELLING IN SIMULATION USING BAYESIAN METHODS

Russell C. H. Cheng

Faculty of Mathematical Studies
University of Southampton
Highfield
Southampton, SO17 1BJ
ENGLAND

ABSTRACT

This paper further develops some of the ideas set out by Cheng (1998) for output analysis using Bayesian Markov Chain Monte Carlo (MCMC) techniques, when a regression metamodel is to be fitted to simulation output. The particular situation addressed by Cheng was where there is uncertainty about the number of parameters needed to specify a model. This arises because there may be uncertainty about the number of terms to be included in the regression model to be fitted. The statistically non-standard nature of the problem means that it requires special handling. In this paper we shall use the *derived chain method* suggested by Cheng (1998). However, whereas in that paper the distribution of the response output of interest was assumed to be simply normal, it is typically the case, especially in the study of systems working near their capacity limit, that this distribution is skewed, and moreover the distribution has a support that is effectively bounded below - that is the distribution has a threshold. We describe how the derived MCMC method might be applied in this situation and illustrate with a numerical example involving the simulation of a computer PAD network.

1 INTRODUCTION

We consider the situation where we are attempting to fit a regression metamodel to simulation output, in which there is uncertainty about the number of parameters that there should be in the model. The situation is basically that given by Cheng(1998). For convenience we reintroduce briefly the terminology and notation here.

We suppose that a run of the simulation model of the system of interest yields an output, y and that this response depends on an independent, or design, variable x (this may be vector valued, though in our example we only consider the scalar case). A typical situation might be a queueing system ,where y is the average queue length over the period of the run, and x is the traffic intensity.

Suppose we conduct r simulation runs of the model with each run conducted at a different x value. Let y_j and x_j be the response and design variable values. We now fit a regression metamodel to examine the dependence of y on x :

$$y_j = \eta(x_j, \boldsymbol{\theta}) + z_j, \quad j = 1,\ 2,\ ...,\ r \tag{1}$$

where z is a 'noise' variable modelling the chance variability of the simulation output, and $\eta(x, \boldsymbol{\theta})$ is the regression function of actual interest. We consider the case where we are uncertain about the precise form of $\eta(x, \boldsymbol{\theta})$ and to allow for this we assume that it has the form

$$\eta(x,\ \boldsymbol{\theta}) = \sum_{i=1}^{k} \beta_i f_i(x,\ \boldsymbol{\varphi}_i). \tag{2}$$

where the β_i are unknown coefficients that are to be estimated from the (x_j, y_j), and the $f_i(x, \boldsymbol{\varphi}_i)$ are suitably selected basis functions. If for example they are polynomials, not dependent on unknown parameters $\boldsymbol{\varphi}_i$, then we have a standard polynomial regression problem.

However, we explicitly wish to allow the case where k is unknown. This situation is non-standard for the following reason. Suppose that a particular component, or term, $\beta_i f_i(x, \boldsymbol{\varphi}_i)$, has been included in the model, but is actually not needed. Then the estimate of β_i will be zero or near zero, rendering estimation of the corresponding $\boldsymbol{\varphi}_i$ meaningless. There is numerical instability if we do try to estimate $\boldsymbol{\varphi}_i$ in this situation. A review of this problem is given by Cheng and Traylor (1995).

Cheng (1998) considered a Bayesian Markov Chain Monte Carlo (MCMC) approach to the problem (a good introduction to MCMC is given by Gilks *et al.* 1996; for its application to regression estimation see Young 1977) and proposed a *derived chain method* that is related to

the approach of George and McCulloch (1993), but which is arguably much simpler to apply. However the method was only described for the case where z in (1) is normally distributed. In this paper we consider a more general formulation where z follows a distribution that is skew and which has a support that has bounded left limit.

Section 2 recalls the derived distribution method and sets out how it can be applied in the more general context of skewed, thresholded, errors. In Section 3 we apply the method to an example described by Cheng and Kleijnen (1999) where the problem is to select an appropriate regression metamodel that attempts to quantify how the delay experienced by packets of characters in a computer PAD network depends on the traffic intensity. Some brief conclusions are drawn in the final section regarding the effectiveness of the method.

2 DERIVED MCMC

2.1 Derived posterior distribution

This section follows closely the terminology given in Cheng (1998) but is set out here for convenience. Let $\mathbf{z}$ denote the observations obtained from simulation runs. The distribution of $\mathbf{z}$ depends on a vector of parameters $\boldsymbol{\theta}$ which are unknown and which we wish to estimate.

In the Bayesian formulation, let $\pi(\boldsymbol{\theta})$ denote the density of the prior distribution of $\boldsymbol{\theta}$, and let $\pi(\boldsymbol{\theta}|\mathbf{z})$ denote the density of the posterior conditional distribution given the data $\mathbf{z}$. This latter density can be calculated from Bayes' formula,

$$\pi(\boldsymbol{\theta}|\mathbf{z}) = \frac{p(\mathbf{z}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} p(\mathbf{z}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}. \tag{3}$$

This formula usually assumes that the dimension of $\boldsymbol{\theta}$ is known. We deal with the situation where the precise number of parameters, s say, is not known by initially not explicitly assuming that s has a prior, but instead that there is a maximal model containing s_0 parameters that is definitely adequate. Thus, whatever the 'true' value of s, this value is less than s_0. The prior $\pi(\boldsymbol{\theta})$ and the likelihood $p(\mathbf{z}|\boldsymbol{\theta})$ is well - defined for this maximal model so that the posterior distribution $p(\boldsymbol{\theta}|\mathbf{z})$ can be calculated from (3).We now calculate the following *derived posterior distribution* for s as

$$p_\delta(s = j|\mathbf{z}) = \int_{S_\delta(\boldsymbol{\theta})=j} p(\boldsymbol{\theta}|\mathbf{z})d\boldsymbol{\theta}, \quad j = 1, 2, ..., s_0,$$

where

$$S_\delta(\boldsymbol{\theta}) = \begin{array}{l} number\ of\ components \\ for\ which\ |\theta_i| > \delta\ at\ \boldsymbol{\theta}. \end{array}$$

This derived distribution is very simple to calculate using the MCMC method.

2.2 Markov Chain Monte Carlo

MCMC is a sampling method for calculating the posterior distribution (3) where the denominator is otherwise difficult to obtain. We regard $\boldsymbol{\theta}$ as the state of a certain Markov chain, defined in such a way that the equilibrium distribution is precisely the required posterior distribution with density $\pi(\boldsymbol{\theta} \mid \mathbf{z})$. We then simulate the Markov chain making the simulation run sufficiently long so that equilibrium is reached. At this stage the sample distribution of the observed $\boldsymbol{\theta}'s$ will have converged to have the required density $\pi(\boldsymbol{\theta} \mid \mathbf{z})$.

We follow Cheng (1998) and use the *Metropolis - Hastings* (MH) algorithm to generate the successive states of the chain: $\boldsymbol{\theta}^0$, $\boldsymbol{\theta}^1$, ..., $\boldsymbol{\theta}^t$, ... Here the state, $\boldsymbol{\theta}^{t+1}$, at time point $t+1$ is obtained from the previous state, $\boldsymbol{\theta}^t$, by generating a candidate value $\boldsymbol{\varphi}$ from a candidate distribution with density $q(\boldsymbol{\varphi}|\boldsymbol{\theta}^t)$. The notation indicates the possibility that this distribution may depend on $\boldsymbol{\theta}^t$, however we use the *independence sampler* which is the case where the candidate density does not depend on the current state, so that

$$q(\boldsymbol{\varphi}|\boldsymbol{\theta}) = q(\boldsymbol{\varphi}). \tag{4}$$

The candidate value is only accepted with probability

$$\alpha(\boldsymbol{\theta}^t, \boldsymbol{\varphi}) = \min\left(1, \frac{\pi(\boldsymbol{\varphi}|\mathbf{z})q(\boldsymbol{\theta}^t)}{\pi(\boldsymbol{\theta}^t|\mathbf{z})q(\boldsymbol{\varphi})}\right) \tag{5}$$

when $\boldsymbol{\theta}^{t+1} = \boldsymbol{\varphi}$. Otherwise the state remains unchanged with $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t$.

The formula (5) for the acceptance probability depends on the conditional posterior distribution that we are attempting to evaluate. If we use a *reference prior* for $\pi(\boldsymbol{\theta})$ (that is a prior that remains essentially constant over the region where the likelihood $p(\boldsymbol{\theta}|\mathbf{z})$ is appreciable), then the posterior density is proportional to this likelihood,

$$\pi(\boldsymbol{\theta}|\mathbf{z}) \propto p(\boldsymbol{\theta}|\mathbf{z}),$$

and the acceptance probability, using the independence sampler, reduces to:

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\varphi}) = \min\left(1, \frac{p(\boldsymbol{\varphi}|\mathbf{z})q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{z})q(\boldsymbol{\varphi})}\right), \tag{6}$$

which depends purely on known quantities. The MH algorithm then has the form:

Initialise $\boldsymbol{\theta}^0,\ t := 0$
Repeat
{
  Generate $\boldsymbol{\varphi} \sim q(.),\ U \sim U(0, 1)$
  If $U \leq \alpha(\boldsymbol{\theta}^t, \boldsymbol{\varphi})$ Set $\boldsymbol{\theta}^{t+1} := \boldsymbol{\varphi}$
  Else Set $\boldsymbol{\theta}^{t+1} := \boldsymbol{\theta}^t$
  Set $t := t + 1$
}

where $\alpha(\boldsymbol{\theta}^t, \boldsymbol{\varphi})$ is calculated using (6).

2.3 Regression Metamodelling

We now describe how the MCMC method can be used for regression metamodelling, in the case

$$y_j = \sum_{i=0}^{k} \beta_i f_i(x_j) + z_j, \quad j = 1,\ 2,\ ...,\ r \tag{7}$$

where z has distribution with support $\{z \mid z \geq 0\}$ and has cdf $G(z,\ \mu,\ \sigma)$, where μ is the mean of G, and σ is some measure of the dispersion. We write

$$g(z) = dG(z,\ \mu,\ \sigma)/dz \tag{8}$$

for the density of the distribution of z.

We also assume that $f_0(x) = f_0$ is a known constant, so that $\gamma = \beta_0 f_0$ is an offset or threshold, and $w = \gamma + z$ has distribution with cdf $G(w - \gamma,\ \mu,\ \sigma)$, depending on three parameters, one of which is a threshold. This formulation is different from that of Cheng (1998); however the standard normal model is still included as a special case with $\mu = 0$.

We assume that k is unknown. The 'correct' true k is defined in one of two ways.

In the first definition we assume that the non-zero coefficients comprise the set $\{\beta_i \mid i \in I\}$ with $\beta_j = 0$ for $j \notin I$, and define the true k to be the largest i amongst all $i \in I$. The alternative definition is the subset selection version, where one wishes to identify the set of non-zero coefficients precisely; that is, to find the set $\{\beta_i \mid i \in I\}$.

The derived MCMC method allows either definition of this 'correct' k. The method is described in the next sub-section.

2.4 Derived Chain MCMC Method

The derived chain MCMC method is as follows.

1. We use a locally uniform reference prior and an independence sampler of the form (4), so that α takes the simple form (6).
2. We now assume that k_0 is a known upper bound on unknown true k. (The precise value for k_0 is relatively unimportant. In practice it can be arbitrarily large, the main limitation being that there should be sufficient degrees of freedom left to estimate μ and σ.) The unknown parameters are therefore

   $$\boldsymbol{\theta} = (\beta_0, \beta_1, ..., \beta_{k_0},\ \mu,\ \sigma).$$

   We then run the MCMC simulation using the MH algorithm given previously, using an appropriately selected candidate distribution, $q(\boldsymbol{\theta})$ (to be discussed in the next sub-section).

The MCMC simulation does not identify the correct value of k explicitly. However if the true value of k is less than k_0, then we can expect that for most of the $\boldsymbol{\theta}^t$, the components θ_j^t will be near zero for $j = k + 1,\ k + 2,\ ...,\ k_0$. Thus if we select $\delta > 0$ and consider θ_i to be zero for practical purposes, if $\theta_i < \delta$, then we construct a *derived chain* $\{\tilde{k}^t,\ t = 0,\ 1,\ 2,\ ...\}$ corresponding to $\{\boldsymbol{\theta}^t,\ t = 0,\ 1,\ 2,\ ...\}$ simply by setting $\tilde{k}^t$ equal to the largest i for which

$$\left|\theta_i^t\right| > \delta. \tag{9}$$

The distribution of the values of k in the sequence $\tilde{k}^t$ can thus be used to estimate the posterior distribution of k.

2.5 Candidate Distribution

The selection of the candidate distribution in the above model turns out to be quite critical. The only satisfactory way we have found to date is to use of an accurate estimate of the asymptotic normal distribution of the maximum likelihood (ML) estimates of the parameters (Kendall and Stuart, 1979). Let

$$\boldsymbol{\theta} = (\beta_0,\ \beta_1, ..., \beta_{k_0},\ \mu,\ \sigma)^T.$$

Then, writing $\hat{\boldsymbol{\theta}}$ for the maximum likelihood estimator of $\boldsymbol{\theta}^*$, the true parameter value, we have that

$$\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}^*, \mathbf{V}),$$

where, for standard situations, the asymptotic variance, $\mathbf{V}$, can be approximated by the inverse of the information matrix:

$$\mathbf{V} = -\left(\partial^2 L/\partial \boldsymbol{\theta}^2\right)^{-1} \tag{10}$$

where

$$L = \sum_{j=1}^{r} \ln \left\{ g\left(y_j - \sum_{i=0}^{k_0} \beta_i f_i(x_j),\ \mu,\ \sigma \right) \right\},$$

with $g(.)$, the density given in (8).

Clearly, under the assumptions made, the posterior distribution will tend to this asymptotic distribution. However in finite samples, there is sufficient discrepancy between this and the true finite sample distribution to make application of the Bayesian technique of potential value. In fact, with certain parameters, we found the difference is sufficiently marked for it to be worthwhile to use a skewed candidate distribution, with mean and variance matched to the values of the estimates of those of the asymptotic normal distribution. We shall discuss this more fully in due course.

We illustrate application of the above to computer network queueing model.

3 APPLICATION

3.1 Inverse Gaussian Model

In the application we used the following explicit model where the errors are assumed to have an inverse Gaussian distribution. A big advantage of assuming this form of error model, is that the threshold estimator is normally distributed like any other standard parameter, unlike many other distributions with a threshold, like the gamma or Weibull. We assume that

$$y_j = \sum_{i=0}^{k} \beta_i f_i(x_j) + w(x_j) z_j, \quad j = 1, ..., n, \tag{11}$$

where $z_j \sim IG(\lambda,\ \mu)$ has density

$$g(z \mid \lambda,\ \mu) = \left(\frac{\lambda}{2\pi z^3} \right)^{1/2} \exp\left(-\frac{\lambda(z-\mu)^2}{2z\mu^2} \right), \tag{12}$$

and that the basis functions f_i are orthonormal. In our numerical example, each $f_i(.)$ was a polynomial of degree i. To allow for heteroscedasticity we introduced weights $w(x_j)$. As the example is for illustration only, these were actually estimated directly from the data in a separate preprocessing stage, and were then subsequently regarded as fixed.

ML estimates were obtained using a mongrel optimization method. The observations were taken in the form $y_j = \sum_{i=1}^{k} \beta_i f_i(x_j) + w(x_j) z'_j$ where z'_j has the three parameter density $g(z-\gamma \mid \lambda,\ \mu)$ (so that $\gamma = \beta_0 f_0$). Now for *fixed* $\boldsymbol{\beta}' = (\beta_1,\ \beta_2,\ ...,\ \beta_{k_0})$ we can treat the

$$z'_j(\boldsymbol{\beta}') = [y_j - \sum_{i=1}^{k} \beta_i f_i(x_j)]/w(x_j), \quad j = 1, 2, ..., r \tag{13}$$

as a sample from $IG(z-\gamma, \lambda, \mu)$. Cheng and Amin (1981) give a method of ML estimation from this distribution using Newton-Raphson iterations based on the updating formula:

$$\gamma_0 = z'_{(1)} - (2s^3 \ln r)^{-1} (\bar{z}' - z'_{(1)})^3$$

$$\mu_m = \bar{z}' - \gamma_m,$$

$$\lambda_m = \left\{ r^{-1} \sum_j (z_j - \gamma_m)^{-1}/r - \mu_m^{-1} \right\}^{-1} \tag{14}$$

$$\gamma_{m+1} = \gamma_m + \rho[3 \sum_j (z_j - \gamma_m)^{-1}/r \ +$$

$$\lambda_m \{\mu_m^{-2} - \sum_j (z_j - \gamma_m)^{-2}/r\}]/\{3\mu_m^{-1}\lambda_m^{-1} + 12\lambda_m^{-1}\},$$

for $m = 0, 1, 2, ...$, where we have corrected the typing error in that paper, and replaced the incorrect $\times$ by a solidus. The Newton-Raphson iterations can be unstable at times. We have therefore included a relaxation factor ρ. A value less than unity, say $\rho = 0.5$,ensures more certain convergence. The complete ML estimation method then uses Nelder-Mead simplex search to minimize the loglikelihood

$$L\{\boldsymbol{\beta}', \mathbf{z}'(\boldsymbol{\beta}')\} = \tag{15}$$

$$\sum_{j=1}^{r} \ln \left[g\{z'_j(\boldsymbol{\beta}') - \gamma(\boldsymbol{\beta}') \mid \lambda(\boldsymbol{\beta}'),\ \mu(\boldsymbol{\beta}')\} \right]$$

with respect to $\boldsymbol{\beta}'$, with g as given in (12), and with $z'_j(\boldsymbol{\beta}')$, $\lambda(\boldsymbol{\beta}')$, $\mu(\boldsymbol{\beta}')$ and $\gamma(\boldsymbol{\beta}')$ calculated at each step from (13) and (14).

Once the parameter estimates have been obtained, can we use the MCMC method to obtain estimates of their distribution. The candidate densities can be taken to be approximately normal with mean $\hat{\boldsymbol{\theta}}$ and variance $\mathbf{V} = -\left(\partial^2 L/\partial \boldsymbol{\theta}^2\right)^{-1}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$, where L is calculated from (15). We found that a simple numerical finite difference formula for evaluating this directly from (15) was sufficiently accurate. This saves on significant algebra to explicitly evaluate second derivatives. What did prove worthwhile however

was to transform λ to σ where $\sigma = \lambda^{-1}$ is a more direct measure of variance and then replace the normal candidate distribution for λ by a gamma candidate distribution for σ, with the mean and variance of this gamma distribution equated to the corresponding values of the asymptotic normal distribution of $\hat{\sigma}$.

3.2 PAD Queue Example

Cheng and Kleijnen (1999) describe the fitting of a regression metamodel in an experiment investigating how the delay in processing characters in a PAD queue depends on arrival rate of characters. Observations (scaled as described in Cheng and Kleijnen) from 353 simulations spread over seven selected arrival rate settings are plotted in Figure 1. The behaviour is non-monotonic and requires fitting a high order polynomial before a satisfactory fit is obtained.

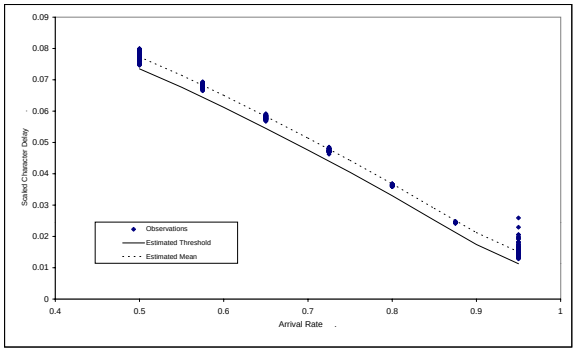

Figure 1: Regression Metamodel Fit for PAD Queue Example

Figure 1 also plots results of fitting the model (11) using the ML estimation procedure of the previous section.

We then used the derived chain MCMC method to estimate the distributions of these parameter estimates. Let $K = k+1$ be the total number of β coefficients. Settings used for the MCMC run were: $T = 50,000.$, $k_0 = 6$, $K_0 = 7$. In Figure 2, the first nine plots give the candidate densities for the parameters β_0, β_1, ..., β_6, σ, μ (smooth curves) together with the histogram of their posterior distributions estimated from the MCMC run. Figure 2 also gives the posterior distribution of $\tilde{K}$, the derived K, using $\delta = 3\tau$ in (9), where τ is the estimate of the average standard deviation of the β estimates. The values $p(\tilde{K} = 1) = 0.040$, $p(\tilde{K} = 4) = 0.878$, $p(\tilde{K} = 5) = 0.055$ and $p(\tilde{K} = 6) = 0.027$, indicate that a degree three polynomial at least is needed, with some evidence that to be on the safe side a polynomial of degree four, possibly even five should be used.

Finally Figure 3, gives a plot of the first 2000 chain states obtained for a selection of the parameters. The plots

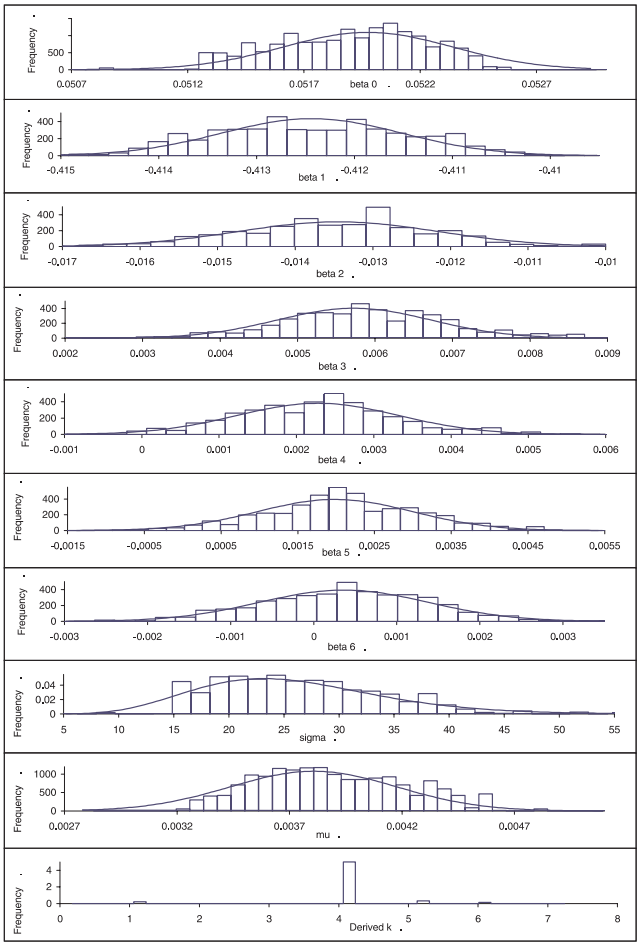

Figure 2: Candidate and Estimated Posterior Distributions of Parameters

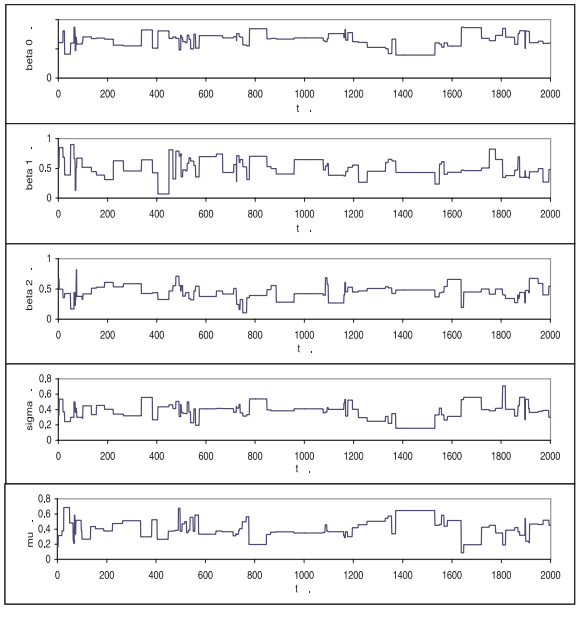

Figure 3: Trace of MCMC States for Selected Parameters

of the trace for the remainder of the run for these parameters and for the remaining parameters were very similar and have not been included. A noticeable feature is how transitions between states only take place irregularly. This accounts for why the histograms in Figure 2 are still somewhat irregular even after $50,000$ iterations. (Part of the difficulty arises because the presence of a threshold means that the values of z_j generated in the runs must be positive if a change of state is to take place.) Even so the histograms are sufficiently stable to give a good indication of the final form of the posterior distributions of the parameters.

4 SUMMARY

We have given a fairly direct way of handling the difficult problem of estimating the unknown number of terms in a regression model, using a simple adaptation of the Bayesian MCMC approach. The method shows some promise, but does require quite careful handling. In particular a careful choice of candidate density seems important. The issue of robustness is therefore of some concern. It would be of some interest to compare the proposed derived MCMC method with the *parametric bootstrap method*; this latter being the other generally used numerically intensive procedure.

REFERENCES

Cheng, R.C.H. and Amin, N.A.K. (1981). Maximum Likelihood Estimation of Paramters in the Inverse Gaussian distribution, with Unknown Origin. *Technometrics*, 23, 257-263.

Cheng, R.C.H. and Traylor, L. (1995). Non-regular maximum likelihood problems (with discussion). *Journal of the Royal Statistical Society, Series B*, 57, 1, 3-44.

Cheng, R.C.H. (1998). Bayesian Model Selection when the Number of Components is Unknown. In *Proceedings of the 1998 Winter Simulation Conference*, eds D.J.Medeiros, E.F. Watson, J.S. Carson and M.S. Manivannan. IEEE, Piscataway, 653-659.

Cheng, R.C.H. and Kleijnen, J.P.C. (1999). Improved Design of Queueing Simulation Experiments with Highly Heteroscedastic Responses, *Operations Research*, To Appear.

George, E.I. and McCulloch, R.E. (1993). Variable Selection via Gibbs sampling. *J. Am. Statist. Ass*. 85, 398-409.

Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

Kendall, M.G. and Stuart, A. (1979). *The Advanced Theory of Statistics. Vol. 2: 4th Edn.* London: Griffin.

Young, A.S. (1977). A Bayesian approach to prediction using polynomials. *Biometrika*, 64, 309-317.

AUTHOR BIOGRAPHY

RUSSELL C. H. CHENG is Professor of Operational Research at the University of Southampton. He has an M.A. and the Diploma in Mathematical Statistics from Cambridge University, England. He obtained his Ph.D. from Bath University. He is Chairman of the U.K. Simulation Society, a Fellow of the Royal Statistical Society, Member of the Operational Research Society. His research interests include: variance reduction methods and parametric estimation methods. He is Joint Editor of the *IMA Journal on Mathematics Applied to Business and Industry*.