Proceedings of the 1998 Winter Simulation Conference
D.J. Medeiros, E.F. Watson, J.S. Carson and M.S. Manivannan, eds.

BAYESIAN MODEL SELECTION WHEN THE NUMBER OF COMPONENTS IS UNKNOWN

Russell C. H. Cheng

Canterbury Business School
The University
Canterbury, Kent CT2 7NF, ENGLAND

ABSTRACT

In simulation modeling and analysis, there are two situations where there is uncertainty about the number of parameters needed to specify a model. The first is in input modeling where real data is being used to fit a finite mixture model and where there is uncertainty about the number of components in the mixture. Secondly, at the output analysis stage, it may be that a regression model is to be fitted to the simulation output, where the number of terms, and hence the number of parameters, is unknown. In statistical terms, such problems are non-standard and require special handling. One way is to use a Bayesian Markov Chain Monte Carlo (MCMC) analysis. Such a method has been suggested by George and McCulloch(1993) using a hierarchical Bayesian model. This method is flexible, but does introduce many additional parameters. This tends to make the modelling look rather complicated. In this paper we adopt a classical Bayesian approach that is essentially equivalent to the George and McCulloch technique, but that has a much less elaborate structure and which renders model interpretation much simpler. The method is illustrated by a regression metamodel example.

1 INTRODUCTION

In computer simulation experiments ever more elaborate models can be used both at the input modelling stage, and at the output analysis stage. We consider the case where we have a parametric model for which there is uncertainty about the number of parameters that there should be in the model. We give two examples.

Cheng, Holland and Hughes (1996) consider modeling the service times of vehicles using a toll booth. A reasonable assumption is that the service times of vehicles of different types, e.g. private cars, light vans, heavy goods vehicles, will have different distributions. A plausible model is therefore to use a *finite mixture* distribution for the service time, X, with density

$$h(x, \boldsymbol{\theta}) = \sum_{i=1}^{k} \beta_i f(x,\ \boldsymbol{\varphi}_i) \tag{1}$$

with $\boldsymbol{\theta} = (\beta_i,\ \boldsymbol{\varphi}_i,\ i = 1, 2,\ ...,\ k)$, where the $\beta_i \geq 0$, $i = 1, 2,\ ...,\ k$, are weights summing to unity

$$\sum_{i=1}^{k} \beta_i = 1,$$

and the $f(x,\ \boldsymbol{\varphi}_i)$ are the densities of the k *components* of $h(x,\ \boldsymbol{\theta})$. For simplicity we have assumed that the component densities all have the same form, though their parameter values can be different.

The vector of parameters, $\boldsymbol{\theta}$, is assumed not known and will have to be estimated from data obtained from the real system. There may well be uncertainty about k itself. Estimation of k is a *non-standard* problem. This is the case that we consider in this paper.

A second example occurs in the output analysis of a simulation experiment. Suppose we conduct r simulation runs of a queueing model with each run conducted at a different traffic intensity. For the jth run, let the traffic intensity be x_j and let y_j be the average customer waiting time, say, obtained from this run. We may fit a regression metamodel to examine the dependence of y on x. Thus we assume

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

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 may well be uncertain about the precise form of $\eta(x, \boldsymbol{\theta})$ in which case we might take it to be similar in form to (1), that is

$$\eta(x,\ \boldsymbol{\theta}) = \sum_{i=1}^{k} \beta_i f_i(x,\ \boldsymbol{\varphi}_i). \tag{3}$$

Here the β_i are simply coefficients and so are not necessarily positive and do not have to sum to unity. 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, the case where k is unknown and where unknown parameters, $\boldsymbol{\varphi}_i$, appear in the basis functions, is non-standard and again this is the case of interest in this paper.

The reason why the case of unknown k is non-standard is as follows. 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. This renders estimation of the corresponding $\boldsymbol{\varphi}_i$ meaningless and we encounter numerical instability if we do try to estimate $\boldsymbol{\varphi}_i$ in this situation. There is a large literature concerning this problem (see Cheng and Traylor 1995, for a review).

One of the earliest attempts, in the regression case, at a Bayesian formulation was made by Young (1977), who obtained limited theoretical results for the normal model. More recent approaches to the problem (George and McCulloch (1993), Green (1995) and Richardson and Green (1997)), have focused on MCMC simulation methods, and have adopted models of variable dimension in the sense that the unknown k is treated explicitly as a parameter, with a prior distribution; and where the main problem is to calculate the posterior distribution of k. These approaches require the formulation of a valid Markov process which includes k as one of the state variables, including a careful definition of the way transitions can take place between the different possible values of k.

The method proposed by George and McCulloch (1993) uses a hierarchical Bayesian method incorporating additional *latent variables* which in effect act as indicator variables of whether particular components should be included or not. The method proposed by Green (1995) and Richardson and Green (1997) uses a reversible jump Markov representation for moving about parameter spaces of different dimensions. Of the two methods, the one proposed by George and McCulloch appears more transparent and easier to implement and interpret. Nevertheless the hierarchical structure is somewhat elaborate.

In this paper we propose a classical Bayesian formulation in which we calculate what we call a *derived posterior distribution* for k. We call the corresponding numerical MCMC method the *derived chain method*. The method is essentially equivalent to George and McCulloch's latent variable method, but is arguably easier to implement and interpret.

In Section 2 we introduce the derived distribution method and compare it with the latent variable method in an elementary example for which a theoretical analysis is possible. In Section 3 we describe the corresponding derived MCMC method, and in Section 4 we apply it to a simple regression problem and also report results for a simulation experiment where the problem is to select an appropriate regression metamodel that attempts to quantify how the delay experienced by packets in a computer PAD network depends on the traffic intensity.

2 BAYESIAN ANALYSIS

2.1 Derived Posterior Distribution

Let $\mathbf{z}$ denote the sample data (in our case these are the observations obtained from simulation runs). The data depends on a vector of parameters $\boldsymbol{\theta}$ which are unknown and which we wish to estimate.

In the Bayesian framework, we start by supposing that there is a prior distribution of possible values for $\boldsymbol{\theta}$. Let $\pi(\boldsymbol{\theta})$ denote the density of this prior. This prior is then updated by incorporating the data to give a posterior distribution with density $\pi(\boldsymbol{\theta}|\mathbf{z})$, that reflects the best information we have about the distribution of $\boldsymbol{\theta}$ given the data $\mathbf{z}$. Our central aim is therefore to calculate this posterior density. By Bayes Theorem this has form

$$\pi(\boldsymbol{\theta}|\mathbf{z}) = \frac{p(\mathbf{z}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int\ p(\mathbf{z}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}. \tag{4}$$

This formula usually assumes that the dimension of $\boldsymbol{\theta}$ is known. We are however interested in the situation where the precise number of parameters, s say, is not known. We deal with this 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 (4). Now let

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

We can now calculate the following *derived posterior distribution* for s :

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

This derived distribution is very simple to calculate using the Markov Chain Monte Carlo method. We shall show how to do this, but we first give an explicit example illustrating why this derived method is essentially the standard Bayesian version of the latent variable method.

2.2 A Theoretical Example

To illustrate the derived posterior distribution method, we consider an example concerning fitting a simple mixture model. Let $U(0,\ b)$ denote the continuous distribution with density

$$\begin{aligned} u(x \mid b) &= b^{-1},\ \ 0 \le x \le b \\ &= 0,\ \ otherwise \end{aligned}$$

Let $\mathbf{x} = (x_1,\ x_2,\ ...,\ x_n)$ be a random sample drawn from $U(0,\ 1)$, but suppose that we do not know this. Instead we assume that

$$x \sim aU(0,1) + (1-a)U(0,b)$$

where b is a known fixed constant greater than unity, to be explicit we might let $b = 2$. The other parameter, a, is not known, but has to be estimated. Adopting the Bayesian approach we suppose a has prior distribution

$$\pi(a) = u(x \mid 1). \tag{5}$$

Thus the correct model is obtained if $a = 1$. Because the $x_i \sim UID(0,\ 1)$ by assumption, the likelihood takes the simple form

$$p(\mathbf{x}|\ a) = [a + (1-a)b^{-1}]^n. \tag{6}$$

Hence

$$\begin{aligned} p(\mathbf{x}) &= \textstyle\int p(\mathbf{x}|\ a)\pi(a)da \\ &= \textstyle\int_0^1 [a + (1-a)b^{-1}]^n da \\ &= [1 - b^{-n-1}]/[(n+1)(1-b^{-1}]. \end{aligned} \tag{7}$$

Using the derived indicator model approach we take the model to be (the correct) $U(0,1)$ if $a > 1-\delta$, where δ is some suitably chosen small number. If we define the indicator variable

$$\begin{aligned} \beta &= 0 \ \ if \ \ 0 \le a < 1-\delta \\ &= 1 \ \ if \ \ 1-\delta \le a \le 1 \end{aligned} \tag{8}$$

we select the correct model with probability

$$p_\delta(\beta = 1 \mid \mathbf{x}) = \int_{1-\delta}^{1} p(\mathbf{x}|\ a)\pi(a)da / p(\mathbf{x}).$$

Using (5), (6), and (7) we find

$$p_\delta(\beta = 1 \mid \mathbf{x}) = \frac{1 - [1 - (1-b^{-1})\delta]^{n+1}}{1 - b^{-n-1}}. \tag{9}$$

This tends to 1 geometrically as $n \to \infty$, for fixed $b > 1$ and any small $\delta > 0$.

We show that the above is a special formulation of the hierarchical approach; in this case we assume

$$a = \gamma v_1 + (1-\gamma)v_2,$$

where $\gamma,\ v_1$ and v_2 are all unknown parameters with carefully selected priors. The hierarchical Bayesian formulation is:

$$\begin{aligned} x \sim p(x \mid a), \quad & a \sim \pi_1(a \mid \gamma, v_1, v_2), \\ & (\gamma, v_1, v_2) \sim \pi_2(\ \gamma, v_1, v_2) \end{aligned}$$

with the second level prior π_2 chosen as follows. The parameter γ is the latent variable with Bernoulli prior

$$\begin{aligned} \gamma &= 0 \ \ with\ probability \ \ 1-p \\ &= 1 \ \ with\ probability \ \ p \end{aligned}$$

The priors of the other two parameters can be chosen more flexibly. We assume, for example, that

$$v_1 \sim U(1-\delta,\ 1), \quad v_2 \sim U(0,\ 1-\delta).$$

We can recover the standard Bayesian formulation

$$x \sim p(x \mid a),\ \ a \sim \pi(a\ )$$

by integrating out the secondary variables

$$\pi(a) = \int \pi_1(a|\gamma, v_1, v_2)\pi_2(\gamma, v_1, v_2)d\gamma dv_1 dv_2.$$

In our case this gives

$$\begin{aligned} \pi(a) &= (1-p)/(1-\delta) \ \ if \ \ 0 \le a < 1-\delta \\ &= p/\delta \qquad\qquad\quad if\ 1-\delta \le a \le 1 \end{aligned} \tag{10}$$

The indicator variable (8) is thus precisely the latent variable in this case and we have

$$p_\delta(\beta = 1 \mid \mathbf{x}) = \frac{p_1}{p_0 + p_1} \tag{11}$$

where

$$p_0 = [(1-p)/(1-\delta)]\{[1 - (1-b^{-1})\delta]^{n+1} - b^{-n-1}(p/\delta)\}$$

and

$$p_1 = (p/\delta)\{1 - [1 - (1-b^{-1})\delta]^{n+1}\}.$$

The $U(0,\ 1)$ prior (5) of our original standard formulation is recovered simply by setting

$$p = \delta.$$

3 NUMERICAL CALCULATION

3.1 Markov Chain Monte Carlo

The posterior distribution (4) is not usually obtainable in closed form because we cannot easily calculate the denominator, even by classical numerical quadrature. The MCMC method is a sampling method for overcoming this problem. 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 $\pi(\boldsymbol{\theta} \mid \mathbf{z})$. If the Markov Chain can then be simulated and the simulation run made sufficiently long so that equilibrium is reached, then the sample distribution of the observed $\boldsymbol{\theta}'s$ will converge to the required form with density $\pi(\boldsymbol{\theta} \mid \mathbf{z})$.

Let $\boldsymbol{\theta}^0$, $\boldsymbol{\theta}^1$, ..., $\boldsymbol{\theta}^t$, ... denote the successive states of the chain. The following general Monte Carlo method is known as the *Metropolis - Hastings* (MH) algorithm.

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 $\boldsymbol{\theta}^t$. The value is only accepted with probability

$$\alpha(\boldsymbol{\theta}^t, \boldsymbol{\varphi}) = \min\left(1,\ \frac{\pi(\boldsymbol{\varphi}|\mathbf{z})q(\boldsymbol{\theta}^t|\boldsymbol{\varphi})}{\pi(\boldsymbol{\theta}^t|\mathbf{z})q(\boldsymbol{\varphi}|\boldsymbol{\theta}^t)}\right) \qquad (12)$$

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

The MH algorithm thus has the form:

Initialise $\boldsymbol{\theta}^0,\ t := 0$

Repeat

{

Generate $\boldsymbol{\varphi}$ ~ $q(.|\boldsymbol{\theta}^t),\ U$ ~ $U(0,1)$

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 $t := t+1$

}

In theory there is considerable flexibility in the choice of the candidate density $q(\boldsymbol{\varphi}|\boldsymbol{\theta})$. We shall use the so-called *independence sampler* which is the case where the candidate density does not depend on the current state, so that

$$q(\boldsymbol{\varphi}|\boldsymbol{\theta}) = q(\boldsymbol{\varphi}). \qquad (13)$$

In the case where there is great prior uncertainty about the value of $\boldsymbol{\theta}$ it is usual to 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. The posterior density is then proportional to the likelihood,

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

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

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\varphi}) = \min\left(1,\ \frac{p(\boldsymbol{\varphi}|\mathbf{z})q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{z})q(\boldsymbol{\varphi})}\right). \qquad (14)$$

3.2 Regression Metamodelling

We now describe the MCMC method for regression metamodelling, for the case

$$y_j = \sum_{i=1}^{k} \beta_i f_i(x_j) + z_j,\ \ j = 1,\ 2,\ ...,\ r \qquad (15)$$

where z has distribution with cdf $F(\cdot,\ \mu,\ \sigma)$, where μ is the mean of F, and σ is some measure of the dispersion.

In this formulation we have absorbed the usual constant term of the right-hand side into the distribution of z. Thus the standard normal model

$$y = \sum_{i=0}^{k} \beta_i f_i(x) + z, \quad z \tilde{}\ N(0,\ \sigma^2)$$

is included in the formulation (15) if we set

$$\mu = \beta_0 f_0(x), \qquad (16)$$

provided $f_0(x)$ is a constant independent of x. The formulation also allows non-normal, typically skew, errors. Lack of space prevents discussion of this possibility further here.

We now make the key assumption that k is unknown. To clarify the discussion, we define more precisely our interpretation of what constitutes a 'correct' true k.

One definition is to 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 $i \in I$. Thus for example if β_1 and β_3 are non zero, and $\beta_2 = \beta_j = 0$, for all $j \geq 4$, then $I = \{1, 3\}$ and $k = 3$.

An 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 $\{\beta_i \mid i \in I\}$. In our example we would therefore wish to identify I as being the set $\{1, 3\}$. Our method will handle either definition of this 'correct' k.

The main requirement in identifying k, is that the Markov process should have a stationary distribution which possesses a high posterior probability for the correct value of k. The model being fitted actually makes this a rather subtle requirement. The critical difficulty is the hierarchical nature of the model, in the sense that if the correct model has $k = k_T$, then *any* model with $k > k_T$ will be equally good if not better. An undemanding formulation might not exhibit a preference for $k = k_T$ over $k > k_T$ and consequently would then not necessarily yield a posterior probability for k_T that was larger than for any $k > k_T$. If an MCMC Bayesian method is to correctly identify k whilst allowing k to vary, then it must include a mechanism for preferring smaller k to larger k, provided the fit is adequate. Without such a preference, the MCMC will not

necessarily produce a posterior distribution with a large probability located at the correct k.

George and McCulloch (1993) handle this difficulty by proposing a hierarchical framework. However their hierarchical definition uses a fairly elaborate distributional structure. A more general methodology has been developed by Green (1995) and Richardson and Green (1997) using a reversible jump process to model state spaces with variable dimension. It is not immediately clear that the process will necessarily identify the correct k. The above methods seem quite elaborate to set up involving a large number of parameters and requiring some sophistication in the interpretation of results. In contrast the derived posterior distribution is easily calculated. The method is described in the next sub-section.

3.3 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 (13), so that α takes the simple form (14).

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

$$\boldsymbol{\theta} = (\beta_1, \beta_2, ..., \beta_{k_0},\ \mu,\ \sigma).$$

We assume that these have prior distribution with density

$$\pi(\boldsymbol{\theta}) = \pi(\beta_1)\pi(\beta_2)...\pi(\beta_{k_0})\pi(\mu)\pi(\sigma).$$

The data $\mathbf{z} = (z_1,\ z_2,\ ...,\ z_r)^T$ in the Bayes formula (4) is calculated from (15) using

$$z_j = y_j - \sum_{i=1}^{k_0} \beta_i f_i(x_j),\ \ j = 1,\ 2,\ ...,\ r.$$

For an appropriately selected candidate distribution, $q(\boldsymbol{\theta})$ (to be discussed in the next sub-section), the MH algorithm reduces to

Initialise $\boldsymbol{\theta}^0,\ t := 0$
Repeat
{
Generate $\boldsymbol{\varphi}$ ~ $q(.),\ U$ ~ $U(0,1)$
If $U \leq \min\left(1,\ \left[p(\boldsymbol{\varphi}|\mathbf{z})q(\boldsymbol{\theta}^t)\right] / \left[p(\boldsymbol{\theta}^t|\mathbf{z})q(\boldsymbol{\varphi})\right]\right)$
Set $\boldsymbol{\theta}^{t+1} := \boldsymbol{\varphi}$
Else Set $\boldsymbol{\theta}^{t+1} := \boldsymbol{\theta}^t$
Set $t := t + 1$
}

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 $\boldsymbol{\theta}_j^t$ will be near zero for $j = k+1,\ k+2,\ ...,\ k_0$. Thus if we select $\delta > 0$ and consider $\boldsymbol{\theta}_i$ to be zero for practical purposes, if $\boldsymbol{\theta}_i < \delta$, 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

$$\left|\boldsymbol{\theta}_i^t\right| > \delta. \tag{17}$$

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

3.4 Candidate Distribution

For the candidate distribution in our proposed methods we recommend use of an estimate of the asymptotic normal distribution of the maximum likelihood estimates of the parameters (Kendall and Stuart, 1979). The asymptotic distribution has to be estimated because it depends on the unknown parameters themselves, which therefore have to be estimated. As we are using an independence sample, the candidate distribution is obtained prior to the simulation and is not altered subsequently. The main requirement is the form of the candidate should have the characteristics known to be desirable (see Gilks et al 1996, page 10). In the examples, below we use a method based on least-squares which is precisely the maximum likelihood method when the model is normal, and is only approximately so when the model is not normal.

We illustrate application of the above to two simple models.

4 APPLICATIONS

4.1 Normal Model

To be explicit we suppose the data has the form:

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

where $z_j \sim N(0, \sigma^2)$, and that the basis functions f_i are orthonormal, that is:

$$\begin{array}{c} \sum_{l=1}^{n} f_i(x_l) = 0\ \ i = 1, 2, ..., k \\ \sum_{l=1}^{n} f_i(x_l) f_j(x_l) = \delta_{ij}\ \ i,\ j = 0, 2, ..., k. \end{array}$$

In our numerical example, each $f_i(.)$ was a polynomial of degree i. In matrix form

$$\mathbf{y} = \mathbf{A}\boldsymbol{\beta} + z,$$

where $\mathbf{A^T A} = \mathbf{I}$, so that the estimates for $\boldsymbol{\beta}$ and σ^2 are

$$\hat{\boldsymbol{\beta}} = \mathbf{A^T y}$$

and

$$s^2 = \nu^{-1}(\mathbf{y} - \mathbf{A}\hat{\boldsymbol{\beta}})^{\mathbf{T}}(\mathbf{y} - \mathbf{A}\hat{\boldsymbol{\beta}})$$

with distributions

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2\mathbf{I}),$$

and

$$\nu s^2 \sim \sigma^2 \chi^2_\nu,$$

where $\nu = n - k$ (see for example, Box and Tiao, 1973). For the candidate distribution $q_k(\boldsymbol{\theta})$ we therefore assume

$$\boldsymbol{\beta} \sim N(\hat{\boldsymbol{\beta}}, \sigma^2\mathbf{I}) \quad and \quad \sigma^2 \sim s^2\chi^2_\nu/\nu.$$

In this simple model we actually know the correct result. For the correct k :

$$\boldsymbol{\beta} \sim N(\hat{\boldsymbol{\beta}}, \sigma^2\mathbf{I}) \quad and \quad \sigma^2 \sim \nu s^2 \chi^{-2}_\nu. \tag{18}$$

Thus the candidate distribution is precisely the posterior for $\boldsymbol{\beta}$, but not for σ^2. A possible alternative is to use this correct version (18) as the candidate distribution, though we do not follow this possibility here.

Figure 1 gives results for the model:

$$\left.\begin{array}{l} y_{jh} = 3x_j + z_{jh}, \\ x_j \ \ = j \end{array}\right\} \ j = 1, ..., 5, \quad h = 1, 2, ..., 10,$$

where

$$z_{jh} \sim N(1,\ 1^2).$$

In terms of orthonormal basis functions this gives

$$\beta_0 = 70.71 \quad \beta_1 = 30.0 \quad \beta_j = 0, \ \ j \geq 2.$$

The MCMC used $T = 50,000$. Let $K = k + 1$ be the total number of β coefficients, with $k_0 = 4$, $K_0 = 5$. The correct value is $K = 2$, so that the correct distribution for K is $p_1 = 0$, $p_2 = 1$, $p_3 = p_4 = p_5 = 0$. In Figure 1, the first six plots give the candidate densities for the parameters β_0, β_1, ..., β_4, σ (smooth curves) together with the histogram of their posterior distributions estimated from the MCMC run. The last two graphs give the posterior distributions of K and the derived version, $\tilde{K}$. For the unadjusted posterior distribution of K, the probability is incorrectly concentrated at $K = 5$. In contrast the posterior distribution of the derived $\tilde{K}$ process with $\delta = 3s$ in (17), yielded a value for $p(\tilde{K} = 2) = 0.9978$ that is very close to unity, the next highest values being already small: $p(\tilde{K} = 4) = 0.0010$ and $p(\tilde{K} = 5) = 0.0009$.

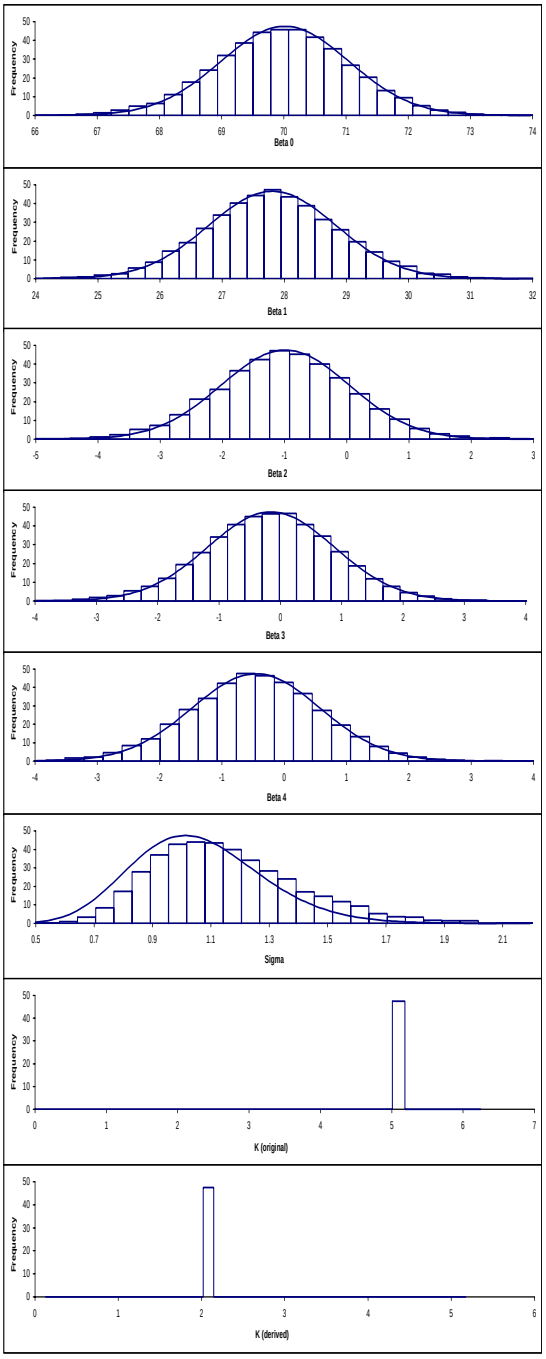

Figure 1: Conditional and Posterior Distributions of Parameters in the Regression Model

4.2 PAD Queue Example

Cheng and Kleijnen (1997) describe the fitting of a regression metamodel in an experiment investigating how the delay in processing characters in a PAD queue depends

on traffic intensity. The behaviour is non-monotonic and required fitting a polynomial of order five or six before a satisfactory fit is obtained. The above Bayesian derived chain method the posterior distribution for the degree of the polynomial gave very similar results. Lack of space prevents a full discussion of this example here, but it is hoped to report on this example elsewhere.

5 SUMMARY

We have given what we consider to be a very simple and direct way of handling the difficult problem of estimating the unknown number of components of a finite mixture model or the unknown number of terms in a regression model, using a simple adaptation of the Bayesian MCMC approach. In limited experiments, two of which are reported here, the proposed method appears quite robust. For example, very similar results were obtained in the above normal regrssion model when the proposed the candidate distributions for the coefficients β have twice the variance of the estimated asymptotic distribution used in the MCMC simulation of Figure 1.

REFERENCES

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., Holland, W. and Hughes, N.A. (1996). Selection of input models using bootstrap good-ness-of-fit. In *Proceedings of the 1996 Winter Simulation Conference*, eds J.M. Charnes, D.J. Morrice, D.T. Brunner and J.J. Swain. IEEE, Piscataway, 317-322.

Cheng, R.C.H. and Kleijnen, J.P.C. (1997). Improved Design of Queueing Simulation Experiments with Highly Heteroscedastic Responses, *Operations Research*, To Appear.

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

Box, G.E.P. and Tiao, G.C. (1973). *Bayesian inference in statistical analysis*. New York: wiley.

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.

Green, P.J. (1995). Reversible jump Markov chain monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711-732.

Richardson, S. and Green, P.J. (1997) On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Statist. Soc*, B, **59**. pp 000-000.

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

AUTHOR BIOGRAPHIES

RUSSELL C. H. CHENG is Professor of Operational Research at the University of Kent at Canterbury. 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*, and an Associate Editor for *Management Science*.