Proceedings of the 2003 Winter Simulation Conference
S. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, eds.

PRIOR AND CANDIDATE MODELS IN THE BAYESIAN ANALYSIS OF FINITE MIXTURES

Russell C.H. Cheng
Christine S.M. Currie

Faculty of Mathematical Studies
University of Southampton
Southampton, SO17 1BJ, U.K.

ABSTRACT

This paper discusses the problem of fitting mixture models to input data. When an input stream is an amalgam of data from different sources then such mixture models must be used if the true nature of the data is to be properly represented. A key problem is then to identify the different components of such a mixture, and in particular to determine how many components there are. This is known to be a non-regular/non-standard problem in the statistical sense and is technically notoriously difficult to handle properly using classical inferential methods. We discuss a Bayesian approach and show that there is a theoretical basis why this approach might overcome the problem. We describe the Bayesian approach explicitly and give examples showing its application.

1 INTRODUCTION

Simulation modelling often requires streams of random variables that represent input quantities that are subject to random variations. For example in a queueing model streams of random interarrival times and customer service times are required. Input modelling aims to identify appropriate probability distributions for characterising the behaviour of such variables. There is a significant literature on this topic. A good general reference is Law and Kelton (2000). Much of the focus is on relatively simple situations where input random variables are independently and identically distributed and drawn from well-known distributions such as the normal, lognormal, gamma or Weibull. In the simulation context, two generalisations have also been studied in some detail, namely: (i) where the random variables are multivariate, and (ii) where they are correlated. See for example Nelson and Yamnitsky (1998), Deler and Nelson (2001) and Ghosh and Henderson (2001).

A third generalisation has not been so well discussed where the random variables come from so-called finite mixture distributions. The purpose of this article is to discuss such distributions and their modelling. We discuss briefly their role in simulation and the kind of situation where they are of use. However the bulk of the paper addresses the problem of fitting an appropriate mixture distribution to existing data.

In Section 2 we give a formal statement of the problem of fitting a finite mixture model and show that it is a non-standard one in a precise statistical sense. We discuss why this non-standardness gives rise to difficult theoretical as well as practical issues. In Section 3 we discuss possible solution methods. In particular we concentrate on a Bayesian approach using a computer intensive method. A Markov chain Monte Carlo method could be used with this Bayesian approach, however we have found that an importance sampling procedure is attractive and this is the method described in this article.

There are some significant practical problems arising from the underlying features and characteristics of the problem. It is the authors' contention that any robust and reliable procedure for fitting a mixture model to data, must explicitly address these issues. In Section 4 we discuss these problems and discuss some practical aspects of the Bayesian approach.

Section 5 gives two numerical examples.

2 THE ESTIMATION PROBLEM

A *continuous finite mixture model* is defined as a distribution whose probability density function takes the form

$$h(x) = \sum_{i=1}^{k} \alpha_i f_i(x \mid \theta_i) \tag{1}$$

where $\alpha_i > 0,\ i = 1, 2, ..., k$ and

$$\sum_{i=1}^{k} \alpha_i = 1 \tag{2}$$

are the weights of the components whose individual densities are $f_i(x \mid \theta_i)$ $i = 1, 2, ..., k$. This notation allows for the component densities to be different but it is more usual to assume that they all take the same form, i.e.

$$f_i(x \mid \theta_i) = f(x \mid \theta_i),\ i = 1, 2, ..., k$$

and this will be assumed for the remainder of the paper. A typical example is where the components are normal

$$f(x \mid \theta_i) = \left(\frac{\tau_i}{2\pi}\right)^{1/2} \exp\left(-\frac{\tau_i}{2}(x - \mu_i)^2\right) \tag{3}$$

where

$$\theta_i = (\mu_i, \tau_i)^T .$$

Note that we follow the convention of many Bayesian formulations in using the parameterization τ rather than the commonly used variance parameter σ^2 . In this model it is assumed that none of the θ_i are known and moreover the number of components *k* is also not known. This is a non-regular situation and there is a large literature discussing the problem. See Cheng (1998) for a discussion of the issues that are of particular concern in this paper. When the number of components is not known, the main problem is that standard asymptotic theory does not hold, and it is difficult to construct satisfactory statistical tests to identify the correct number of components. In what follows we shall assume that the true (but unknown) number of components is denoted by k^*, and the unknown true parameters associated with the individual components are denoted by θ_i^*, $i = 1, 2, ..., k^*$.

3 A BAYESIAN FITTING METHOD

Certain of the methods that have been proposed for fitting mixture models are theoretically interesting but seem rather elaborate to implement or require careful, sometimes rather subjective, selection of key critical values needed by the methodology. An example is the so-called sieve method which requires a sequence of parameter regions of increasing size to be selected as sample size increases. See Barron, Schervish and Wasserman (1999). An exception is the sequential method described by Hsu, Walker and Ogren (1986). This latter method is one of the easiest to understand and implement and deserves to be better known.

We shall consider just one Bayesian method. It combines a simplicity of approach with a clear rationale that is easily understood.

We shall for simplicity assume that, though k^* is unknown, we can specify an upper bound, *K*, for which we know that

$$k^* < K .$$

Bayesian analysis is well described in Gilks, Richardson and Spiegelhalter (1996), especially the MCMC approach. The basic idea of Bayesian estimation is as follows. We suppose that a prior distribution for the unknown parameters can be specified, which we write in the form

$$\pi(\theta^{(k)} \mid k)\pi(k), \quad k = 1, 2, ..., K \tag{4}$$

where

$$\pi(k), \quad k = 1, 2, ..., K \tag{5}$$

is the prior distribution for *k* and, where for given *k*, $\pi(\psi^{(k)} \mid k)$ is the conditional prior density, given *k*, of the component parameters

$$\psi^{(k)} = (\theta_1,\ \theta_2,\ ...,\ \theta_k, \alpha_1, \alpha_2, ..., \alpha_k), \tag{6}$$

The posterior distribution, when fitting the model (1) to a sample $x = (x_1, x_2, ..., x_n)$, is then given by

$$p(\psi^{(k)}, k \mid x) = \frac{p(x \mid \psi^{(k)}, k)\pi(\psi^{(k)} \mid k)\pi(k)}{\sum_{k=1}^{K} \pi(k) \int p(x \mid \psi^{(k)}, k)\pi(\psi^{(k)} \mid k) d\psi^{(k)}}, \quad k = 1, 2, ..., K \tag{7}$$

The expression $p(x \mid \psi^{(k)}, k)$ is the likelihood corresponding to the mixture model with *k* components.

The major difficulty in determining $p(\psi^{(k)}, k \mid x)$ is evaluating the denominator in the above expression (7). Markov chain Monte Carlo (MCMC), is the most popular current method of constructing posterior distributions that does not require explicit evaluation of the denominator. However it is not easy to implement in the current situation because it requires random moves between different *k* values, and the form these moves should take is not straightforward to identify. The reversible jump method for doing this, described by Green (1995) and by Richardson and Green (1997), has received much recent attention, but seems quite elaborate to implement. A more attractive approach would seem to be that of George and McCulloch (1993) where indicator variables are used. A very simple variation of this approach using an embedded MCMC, and

that does not need the use of indicator variables, is described by Cheng (1998).

An alternative to MCMC is importance sampling. The general consensus seems to be that importance sampling is less robust than MCMC when the form of the posterior is not all that well known in advance. However in the case of mixture models we have found that it is much more easy to implement than MCMC. Our suggested procedure is as follows:

Conditional on k we use numerical search to find

$$\tilde{\psi}^{(k)} = \arg\max_{\theta^{(k)}}[p(x \mid \psi^{(k)}, k)\pi(\psi^{(k)} \mid k)] \tag{8}$$

for each $k = 1, 2, ..., K$. For given k, the problem is regular, and as the sample size n tends to infinity the asymptotic distribution of $\tilde{\psi}^{(k)}$ is normal with variance given by the inverse of the information matrix at $\tilde{\psi}^{(k)}$. Interpreting this result from the Bayesian viewpoint yields a limiting posterior distribution for $\psi^{(k)}$ that is multivariate normal, namely

$$q(\psi^{(k)} \mid k) = \Phi(\psi^{(k)} \mid \tilde{\psi}^{(k)}, \Xi^{(k)}) \tag{9}$$

where $\Phi(\psi^{(k)} \mid \tilde{\psi}^{(k)}, \Xi^{(k)})$ is the (degenerate) multivariate normal density with mean $\tilde{\psi}^{(k)}$ and variance $\Xi^{(k)}$. This distribution is degenerate because condition (2) has to be satisfied.

For the importance sampling procedure we can now construct a candidate distribution as follows. The candidate distribution for the number of components is the uniform distribution

$$q(k) = K^{-1}, \; k = 1, 2, ..., K \tag{10}$$

The complete candidate distribution is simply

$$q(\psi, k) = q(k)q(\psi^{(k)} \mid k) = K^{-1}\Phi(\psi^{(k)} \mid \tilde{\psi}^{(k)}, \Xi^{(k)}) .$$

The importance sampling is now easily implemented. A sample, of size m, of values $(k_j, \psi_j^{(k_j)}), \; j = 1, 2, ..., m$, is drawn from the candidate distribution $q(\psi^{(k)} \mid k)q(k)$ as given in (9) and (10). The posterior distribution sample is then given by

$$p(\psi^{(k_j)} \mid x) = \frac{p(x \mid \psi^{(k_j)}, k_j)r(\psi^{(k_j)}, k_j)}{\sum_{j=1}^{m} p(x \mid \psi^{(k_j)}, k_j)r(\psi^{(k_j)}, k_j)}, \tag{11}$$

$$j = 1, 2, ..., m$$

where

$$r(\psi^{(k_j)}, k_j) = \frac{\pi(\psi^{(k_j)} \mid k_j)\pi(k_j)}{q(\psi^{(k_j)} \mid k_j)q(k_j)}. \tag{12}$$

An attractive feature of the method is that the posterior distribution sample is a *random sample*. Moreover if the candidate distribution is similar in shape and location to that of the posterior distribution then the sample values (11) will be fairly constant and so will accurately reproduce the shape of the posterior distribution without need for the sample size m to be all that large. In contrast the MCMC procedure gives rise to a correlated sequence with convergence properties that can be hard to establish.

4 PRACTICAL IMPLEMENTATION

A central problem in fitting mixture models, when using a classical likelihood approach, is how to obtain the "best" fit. This problem occurs even when the number of components is given. This problem is actually inherent in the problem rather than being a mere artefact of estimation method. The same problem is thus encountered with the Bayesian approach described in the previous section and manifests itself in the following way.

Note first that, in order to identify a candidate distribution that is a good approximation of the posterior distribution, we need to find the maximum of this posterior distribution so that the candidate distribution can have the bulk of its probability located in the neighborhood of this maximum point. The key calculation is thus that of (8).

In evaluating (8), the Bayesian approach possesses a distinct theoretical advantage over classical methods that maximize the likelihood or any modified form of likelihood. The problem with a mixture model is that it possesses unusual flexibility and this can result in spurious 'good' fits being obtained. A simple example typifies the problem. Suppose that the data is actually drawn from a one-component model, but that we fit a two-component model. Then only one component is needed to give a satisfactory fit. This means that the second component is available to be fitted to any minor departure of the data, not accounted for by the main fit. To take the most extreme case, the spare component can even be fitted any *one* of the individual observations. This arbitrariness is reflected in the likelihood function which is *always* multimodal with infinite spikes on the boundary of the parameter space corresponding to degenerate delta function components positioned on individual data points.

For example if we place a normal component (3) on the datum point x_1 then this will contribute the term

$$\ln[\alpha_1 f(x_1 \mid \mu_1, \tau_1)] = \frac{1}{2}\ln(\frac{\alpha_1^2\tau_1}{2\pi}) - \frac{\tau_1}{2}(x_1 - \mu_1)^2$$

to the likelihood function. It is easily seen that there are paths in the parameter space with $\mu_1 \to x_1$ and $\tau_1 \to \infty$ for which $\ln[\alpha_1 f(x_1 \mid \mu_1, \tau_1)] \to \infty$.

There are various ways of overcoming the problem, but the Bayesian approach handles the difficulty in a natural way. Whereas in the likelihood approach, preference is given to parameter values where the likelihood is large, in the Bayesian context it is actually the probability content that counts. Thus though a narrow component might in principle be located at a datum point, the probability content of the component will be determined by the weighting of the component; and this will be small.

The above consideration indicates why Bayesian importance sampling might prove preferable to classical maximum likelihood methods. However the problem of satifactorily evaluating (8) is still not straightforward.

We suggest that a good approach for determining the best fit as given in (8), is to use a variant of the collapsing method suggested by Sahu and Cheng (2003). The method begins by selecting $K > k^*$ and then fitting a K component model by finding $\widetilde{\psi}^{(K)}$ as in (8).

When $K > k^*$ the arbitrariness in how components might be fitted in different ways to a sample, is most marked and problematic. This is because there is an ample number of components to account for all the main features of the sample, which then leaves a freedom, but arbitrariness, in how the additional components might be deployed. At least two different phenomena can arise.

One possibility is where two or more model components combine to account for just one true component. This is not actually of great concern. It is merely a manifestation of the redundancy in the fit.

The other possibility is that model components will be fitted to minor, spurious or random, features of the sample. Again this is will not usually be of consequence, as such model components will not be significant, statistically speaking.

The important point to note is that such arbitrariness in the fit is not of real consequence provided that genuine features of the data are captured in the fit.

Assuming that a satisfactory fit has been found with the K-component model, we then progressively reduce the number of components in a step by step manner. The Sahu and Cheng method does this by identifying the pair of components that are closest together according to a certain information-theoretic distance, and combining this pair of components. This method is fairly robust, but it may not corectly handle the situation where two components are close according to the distance measure, but both are nevertheless needed to model an important feature of the data. For example, if the components are symmetric, a skewed mode might need two such components to capture its character. We therefore suggest a procedure that is still stepwise, but which is more cautious, and can deal adequately with such a situation.

Suppose we have a k-component model fitted. We then drop each of the k fitted components in turn, and reoptimize as in (8). We then select the best of the resulting (k-1) component models as being the selected (k-1) component model fit.

More elaborate sequential procedures can be adopted but we have not investigated this possibility any further. The procedure described above has been sufficiently robust in all the examples we have studied to date.

Another issue that is important in considering the number of components needed, is the shape of the distribution assumed for the component densities. For example if the components are normal distributions, and so are symmetric, then an unnecessarily, in effect incorrectly, large number of components will be required to represent data that is a mixture of skew components.

5 EXAMPLES

In the examples we used a normal mixture with components as given in (3). For the parameters μ_i a normal prior was used with the sample mean and sample variance as hyper-parameters; for the τ_i we used an exponential prior; for the weights α_i we used a prior uniform Dirichlet distribution.

The first example considers three sets of observed processing times of vehicles at toll booths of one of the Severn Bridge river crossings. This data was originally reported by Griffiths and Williams (1984), and is also described by Cheng, Holland and Hughes (1996). The data are for vehicles grouped into three categories: private cars, light vans, and heavy goods vehicles. We carried out the Bayesian fitting procedure to observations from each group on their own using a normal mixture model; we also applied the fitting procedure with all three groups combined pretending that the groupings were unknown. The frequency plots and fitted mixture distributions are shown in Figure 1, with the calculated posterior distributions for the number of components given in Figure 2. It will be seen that there is clear evidence that both the Private cars and the HGV's have a distribution where the bulk of the observations can be explained by one component, but that there is a, rather distinct, long tail that requires a separate distinct component. The light vans are a more uniform group, and though the data seem slightly skewed, the mixture model fitting indicates that one component is quite sufficient.

When all three groups are combined, then there is a marked skewness that is distinct from the long tail behaviour, and the posterior distribution now indicates the need for two main components to explain this skewness *and* a separate component to handle the separate long tail.

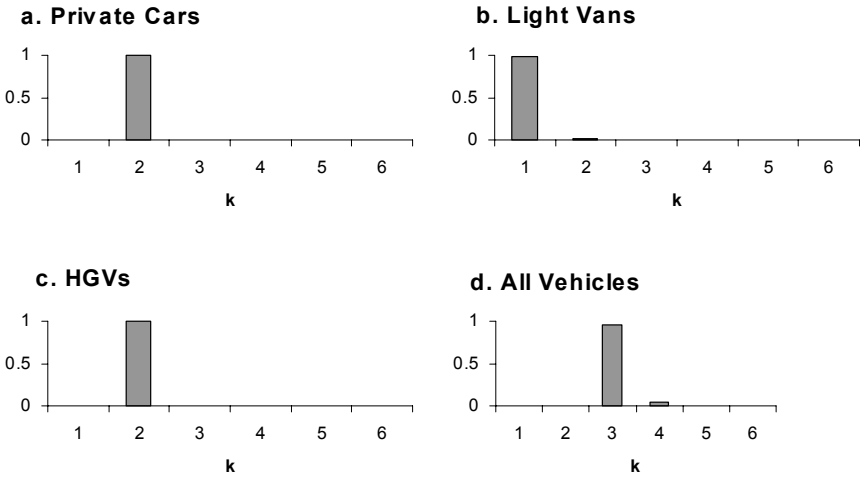

Figure 1: Posterior Distributions for the Number of Components in a Normal Mixture Describing Data from (a) Private Cars, (b) Light Vans, (c) HGVs and (d) All Vehicles

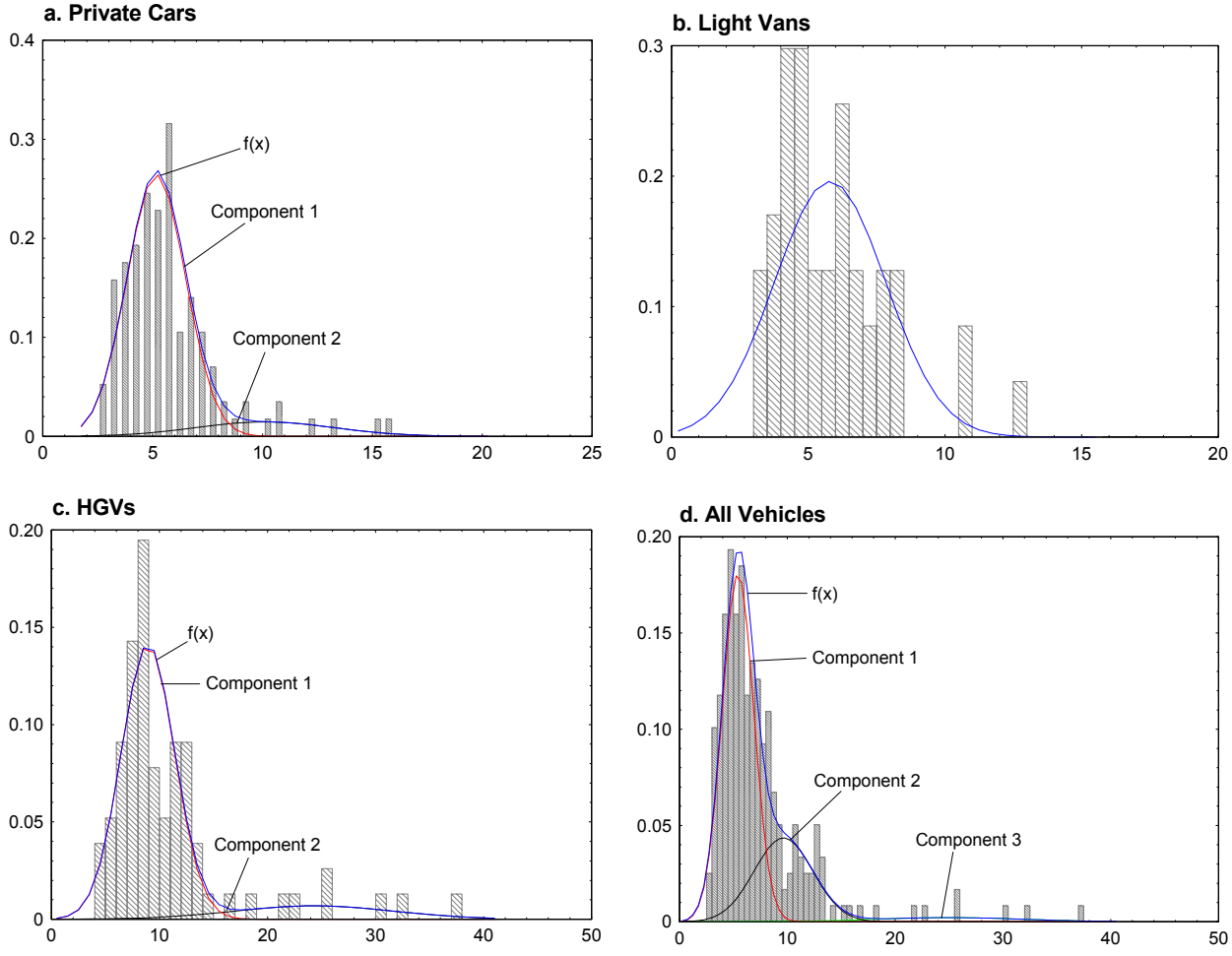

Figure 2: Fitted Mixture Distributions with Individual Components Shown for (a) Private Cars, (b) Light Vans, (c) HGVs, (d) All Vehicles

The second example is taken from observations of headway gaps between vehicles in a separate traffic survey by Cowan (1975). There are theoretical reasons for believing that the distribution in this case is positively skewed. Figure 3 shows the data and Figure 4 gives the posterior distribution of the number of components using a mixture of normal distributions. The spread of probabilities across a number of components is a good indication that the shape of the distribution (normal in this case) used for the individual components is not very appropriate. If we log the data then a very good fit is obtained from just one component as is shown in Figure 5.

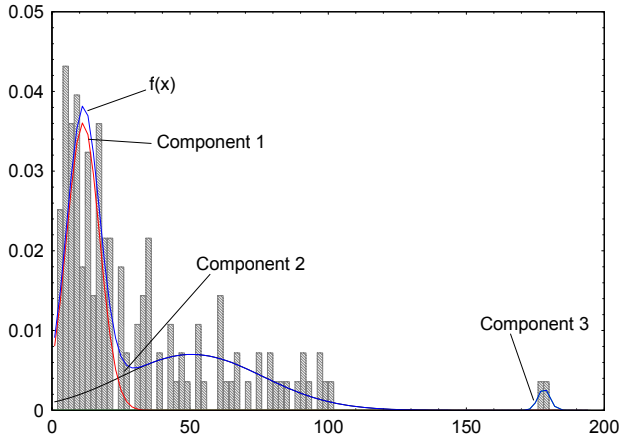

Figure 3: Fitted Mixture Distribution with Individual Components Shown for the Cowan Data

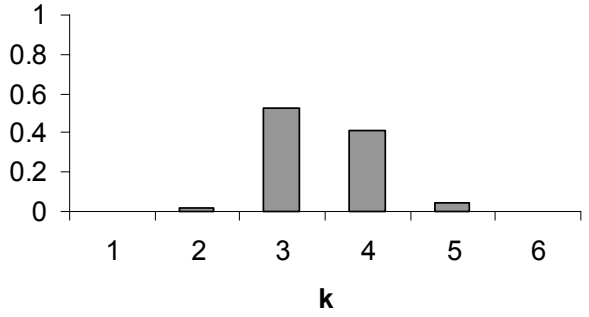

Figure 4: Posterior Distribution for the Number of Components in a Normal Mixture Describing the Cowan Data

6 CONCLUSION

We have tested, fairly extensively, the method for fitting mixture models proposed in this paper, and have found it to be quite robust. The major computational problem is in obtaining the maximum point $\tilde{\psi}^{*(k)}$, for $k = K$, because of the arbitrariness phenomenon discussed in Section 4. However provided K is sufficiently large (values we have used range from 6-10 when the correct k^* is 2-4) the problem does not seem to give rise to any practical difficulties in estimating the correct value for k^*. We have tested this not only with

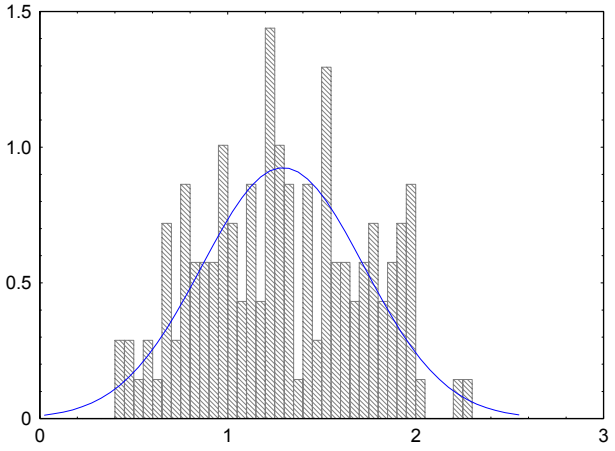

Figure 5: Fitted Normal Distribution to the Log of the Cowan Data

real data but with simulated data where the number of components is known. The method appears to have little difficulty in identifying the correct value for k^*. An Excel spreadsheet version is available at www.maths.soton.ac.uk/staff/Cheng which allows data to be input easily.

REFERENCES

Barron, A., Schervish, M. J. and Wasserman, L. 1999. The consistency of Posterior Distributions in non-parameteric problems. *Annals of Statistics* 120: 536-561.

Cheng, R.C.H. 1998. Bayesian model selection when the number of components is unknown. In *Proceedings of the 1998 Winter Simulation Conference,* ed. D.J.Medeiros, E.F. Watson, J.S. Carson and M.S. Manivannan, 653-659. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Cheng, R.C.H., Holland, W. and Hughes, N.A. 1996. Selection of input models using bootstrap goodness-of-fit. In *Proceedings of the 1996 Winter Simulation Conference*, ed. J.M. Charnes, D.J. Morrice, D.T. Brunner and J.J. Swain, 317-322. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Cowan, R.J. 1975. Useful headway models. *Transportation Research* 9: 371-375.

Deler, B. and Nelson, B.L. 2001. Modeling and Generating Multivariate Time Series with Arbitrary Marginals and Autocorrelation Structures. In *Proceedings of the 2001 Winter Simulation Conference,* ed. B.A. Peters, J.S. Smith, D.J. Medeiros, and M.W. Rohrer, 275-282. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

George, E.I. and McCulloch, R.E. 1993. Variable Selection via Gibbs sampling. *Journal of the American Statistical Asociation* 85: 398-409.

Ghosh, S. and Henderson, S.G. 2001. Chessboard Distributions. In *Proceedings of the 2001 Winter Simulation Conference,* ed. B.A. Peters, J.S. Smith, D.J. Medeiros, and M.W.Rohrer, 385-393. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

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.

Griffiths, J.D. and Williams J.E. 1984. Traffic Studies on the Severn Bridge. *Traffic Engineering and Control* 25: 268-71, 274.

Hsu, Y-S, Walker, J.J. and Ogren, D.E. 1986. A stepwise method for determining the number of component distributions in a mixture. *Mathematical Geology* 18: 153-160.

Law, A.M. & Kelton, W.D. 2000. *Simulation Modeling and Analysis*. New York: McGraw-Hill.

Nelson, B.L. and Yamnitsky. M. 1998. Input Modeling Tools for Complex Problems. In *Proceedings of the 1998 Winter Simulation Conference,* ed. D.J. Medeiros, E.F. Watson, J.S. Carson and M.S. Manivannan, 105-112. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Richardson, S. and Green, P.J. 1997. On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society B* 59: 473-484.

Sahu, S.K. and Cheng, R.C.H. 2003. A fast distance-based approach for determining the number of components in mixtures. *The Canadian Journal of Statistics* 31: 3-22.

AUTHOR BIOGRAPHIES

RUSSELL C. H. CHENG is Professor, Head of Operational Research, and Deputy Dean of the Faculty of Mathematical Studies 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 a former Chairman of the U.K. Simulation Society, a Fellow of the Royal Statistical Society, and Member of the Operational Research Society. His research interests include: variance reduction methods and parametric estimation methods. He was a Joint Editor of the *IMA Journal of Management Mathematics*. His email and web addresses are <R.C.H.Cheng@maths.soton.ac.uk> and <www.maths.soton.ac.uk/staff/Cheng>.

CHRISTINE CURRIE is a PhD student at the University of Southampton. She has an MPhys from Oxford University and an MSc in Operational Research from the University of Southampton. Her research interests include mathematical modeling of epidemics, Bayesian statistics and variance reduction methods. Her email address is: <C.S.M.Currie@maths.soton.ac.uk>