18th World IMACS / MODSIM Congress, Cairns, Australia 13-17 July 2009
http://mssanz.org.au/modsim09

How Reliable is My Reliability Model?

Burrell, D.L. [1], **S. Low Choy** [1] **and K.L. Mengersen** [1]

[1] *School of Mathematics, Queensland University of Technology*
Email: daniel.burrell@student.qut.edu.au

Abstract: In any statistical analysis the adopted model should be chosen to suit the aim of the analysis. An example in rare event modeling might be the description of tail behavior in a distribution. In these situations it may be useful to combine several distinct model choice criteria which emphasize different aspects of model-fit in discriminating amongst models, so as to ensure their suitability to the specific inferential aim.

In this paper we consider the problem of estimating the reliability of a given system. As a concrete example, we consider the problem of assessing the failure characteristics of a certain type of aircraft air-conditioner unit commonly used to control cabin temperatures in a certain type of passenger aircraft used by most airlines around the world. In this case, interest is less focused on the overall fit of a model and more on the goodness-of-fit of the model to the tails of the distribution, since these regions represent the more extreme reliability outcomes.

We consider four statistical distributions that are commonly employed in reliability settings: exponential, gamma, log-normal and the standard (2 parameter) Weibull distribution. These are formulated in a Bayesian setting with standard uninformative priors, and applied to the example problem discussed above.

A combination of two established approaches to model performance assessment is employed. The first is a measure of a model's overall goodness-of-fit, namely the Deviance Information Criterion (DIC). The second forms a class of measures for predictive performance called Posterior Predictive Checks (PPCs), and these can be readily tailored to assess model-fit to specific regions of interest. In our application we show that the choice of model indeed depends on whether emphasis is on best overall fit of a model to the data, or on more tailored assessment. The Gamma and Weibull models provide the best overall fit to the data based on the DIC. However, the upper and lower 10% of the data is predicted better by the log-normal model, a fact that would be overlooked if model selection were based on the DIC alone. The middle part of the data is predicted best by the gamma model.

Compared with the analogous frequentist assessment of model fit, which uses calculations based on point estimates of parameters without carrying their inherent uncertainty through to the final goodness-of-fit statistic, our approach fully incorporates uncertainty in parameter estimates via the posterior distribution.

Keywords: *Reliability, Model discrimination, Bayesian models, Deviance Information Criterion (DIC), Posterior Predictive Checks (PPC), Rare event simulation.*

1. INTRODUCTION

In many statistical analyses, a large part of the work is in formulating models relevant to the inferential aim, and then discriminating between them. For example, in a mechanical reliability setting, interest might be in quantifying the risk of a failure event prior to a scheduled maintenance event. In such a situation, we are not only interested in the high–density region of a failure distribution but also in the lower–density regions which characterize rarer events. It is therefore of interest to consider model comparisons tailored to emphasize local regions of the data space as well as the more common whole-of-distribution comparisions.

Alston et al. (2005) briefly review some popular approaches to model comparison in a Bayesian framework, considering the advantages and disadvantages of each approach in the practical context of mixture models. Three main approaches under a Bayesian framework are identified, namely methods based on separate estimation of the parameters of potential models, methods based on comparative estimation of potential-model parameters and methods based on simultaneous estimation of potential-model parameters. Separate estimation approaches include posterior predictive distributions and posterior predictive checks, Bayes factors and approximations such as the Bayesian Information Criterion (BIC) and Deviance Information Criterion (DIC). Comparative approaches include the use of distance measures such as entropy distance or Kullback-Leibler divergence. Simultaneous methods include reversible jump MCMC and birth and death processes. Further reviews in the area of Bayesian model selection include Spiegelhalter et al. (2002) and Carlin and Chib (1995). All of the approaches discussed in these articles consider the goodness-of-fit of models to the scope of the whole of the data. It is apparent that tailored, local assessment of models is not common: an informal look at twelve beginners-to-advanced Bayesian textbooks, published over the last fifteen years, shows only three books that mention posterior predictive checks (see, Congdon, 2006; Gelman et al. 1996; and Geweke, 2005), only one of which considers, almost as a passing issue, the notion of tailoring model assessment to local regions of the data.

In this paper we consider combining two common methods for model comparison, both of which fall under the realm of separate estimation methods. A standard model assessment approach used in a maximum likelihood setting relies on statistical or information–theoretic criteria that quantify, in some sense, the overall discrepancy of the model in fitting the data. Assessment of model fit to specific, desirable regions, of the data space is rarely considered, as discussed above. We believe that a combination of model goodness–of–fit criteria should be employed to this end. A Bayesian framework is particularly amenable to this task, especially with respect to assessing model goodness–of–fit to certain desired sub-regions of the data space.

We demonstrate one combined approach to assessing model goodness–of–fit and make application to a concrete example taken from a reliability setting. Al-Garni et al. (2006) published an assessment of the failure characteristics of a certain type of air–conditioner cooling–pack used to control cabin temperatures in a common type of passenger aircraft used by most airlines around the world. Their data recorded cumulative failure times measured in flight hours, for both left and right side cooling packs (see Table 1). They considered the failure data at the level of individual components, combining data for each component across 8 different cooling–packs: left and right packs from each of four distinct aircraft. Data from only two of these units is available to us. As a result, we make the assumption that the components form a homogeneous set, and therefore consider system failures to correspond precisely to component failures.

This paper follows a similar pattern of model development, fitting a selection of the models used in the original analysis (Al-Garni et al., 2006). To consider the overall fit of each model we make use of the deviance information criterion (DIC). We also employ some common posterior predictive checks (PPCs) to consider both overall model fit and tail behaviour.

Table 1. Cumulative failure times for left and right air-conditioning/cooling packs. Times are recorded in flight hours.

Left AC Unit			Right AC Unit		
Failure time	Status	Component	Failure time	Status	Component
90.88	F	Water Separator (WS)	63.04	F	WS
111.93	F	WS	112.18	F	Low-limit Valve (LLV)
168.01	F	WS	167.72	F	WS
333.52	F	Heat Exchanger (HE)	269.27	F	WS
604.25	F	WS	604.25	F	WS
688.13	F	WS	688.13	F	Shutoff Valve (SV)

842.17	F	Compressor Discharge (CD)	794	F	WS
948.72	F	WS	1016.6	F	WS
1106.93	F	WS	1393.58	F	WS
1405.68	F	WS	1568.24	F	WS
1445.22	F	Air-cycle machine (ACM)	1603.43	F	Panel Switch (PS)
1568.24	F	WS	1649.14	F	HE
1603.43	F	WS	1688.26	F	WS
1649.14	F	HE	1851.84	F	CD
1796.84	F	WS	2042.74	F	ACM
1804.21	F	WS	2360.54	F	WS
2010.74	F	WS	2378.01	F	SV
2231.63	F	Ram inlet actuator (RIA)	2459.37	F	WS
2378.01	F	WS	2476.17	F	WS
2459.37	F	WS	2607.41	F	WS
2481.57	F	PS	2680.23	F	WS
2607.41	F	WS	2691.87	F	WS
2676.58	F	WS	2765.43	F	Ram Air (RA)
2731.53	F	SV	2812.93	F	WS
2765.43	F	WS	2874.6	F	CD
2814.1	F	HE	3082.42	F	WS
2855.58	F	WS	3212.05	F	HE
2990.27	F	CD	3368.9	F	WS
3046.54	F	WS	3717.82	F	WS
3212.05	F	HE	3901.38	F	WS
3368.9	F	WS	4014.72	F	ACM
3405.62	F	ACM	4098.58	F	WS
3513.53	F	WS	4213.4	F	PS
3888.26	F	WS	43557.17	F	WS
4098.58	F	WS	4810.06	S	--
4423.9	F	LLV			
4578.89	F	WS			
4736.8	F	WS			
4810.06	S	--			

2. METHODS

Al-Garni et al. (2006) use maximum likelihood methods to produce parameter estimates for six commonly used time–to–event models: exponential, gamma, log-normal, 2 and 3 parameter Weibull, as well as a two-component mixture of Weibull densities and a phase-shifted bi-Weibull model. We consider fitting four of these models, as listed for a time–to–failure, t:

(M1) Exponential Model

$$f(t;\lambda) = \lambda \exp\left(-\lambda t\right);$$

(M2) Gamma Model

$$f(t;\alpha,\lambda) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} \exp\left(-\lambda t\right);$$

(M3) Log-Normal Model

$$f(t;\mu,\tau) = \left(\frac{\tau}{2\pi}\right)^{\frac{1}{2}} \frac{1}{t} \exp\left[-\frac{\tau}{2}\left(\log(t)-\mu\right)^2\right];$$

(M4) Weibull Model

$$f(t;\alpha,\lambda') = \alpha\lambda' t^{\alpha-1}\exp\left(-\lambda' t^{\alpha}\right).$$

In the above, we use α to denote shape parameters and λ to denote rate parameters. For identifiability, when there is need to distinguish between parameters coming from different models, the parameters will be subscripted with their associated model labels: M1 through M4. For model 3 we follow the Bayesian conventional notation for a log-normal density, namely for μ the location parameter and τ for the precision (inverse–variance) parameter. In model 4 we employ $\lambda' = \lambda^{\alpha}$, where α and λ are the respective shape and rate parameters of the Weibull density. In all cases, $0 < \alpha < \infty$, $0 < \lambda < \infty$. Also $-\infty < \mu < \infty$, $0 < \tau < \infty$, $0 < \lambda' < \infty$ and $0 < t \leq \infty$.

In order to emphasize the role of the model and corresponding likelihood, we follow advice given by Press (2003) in choosing improper, flat priors for the model parameters. These are implemented on the natural logarithm scale for parameters defined on the non–negative real line and on the identity scale for parameters defined on the entire real line. Of course, in many practical situations substantive prior information may be available and it will be desirable to encode this into appropriate proper prior probability models. This is, indeed, one of the primary practical motivations for the use of the Bayesian approach.

For model comparisons we employ the DIC as a measure of a model's overall goodness-of-fit to data. The DIC is defined as $DIC(\theta) = \overline{D(\theta)} - p_D$, where $\overline{D(\theta)}$ is the expectation of the deviance for the model with parameter vector θ and p_D is a penalty term representing the effective number of parameters in the model (Press, 2003). PPCs are used to consider the goodness-of-fit of a model to particular regions of the support of the posterior. A PPC is based on simulated replicate datasets from the posterior predictive distribution $p(y^{rep} \mid y) = \int_{\theta} p(y^{rep} \mid \theta)\, p(\theta \mid y)\, d\theta$ that correspond to the observed data. A set of simulated observations, $y_i^{(m)}$, $i = 1,\ldots,N;\ m = 1,\ldots,M$ where N is the size of the sample data, and M is the size of the simulation data, is drawn from the posterior predictive distribution. Various summary or discrepancy statistics, $t(y^{\{(m)\}})$, are then evaluated for each m. For example, $t(y^{\{(m)\}}) = y_{\min}$ indicates that for each of the M simulated datasets of size N, the minimum observation is recorded to produce an empirical distribution of posterior predictive minimum observations, $p(t(y^{\{(m)\}}) \mid y)$. This is then compared to the observed minimum observation to give a sense of how well the model predicts the low extreme of the observed data. More generally, any statistic $t(y^{\{(m)\}})$ may be employed and the posterior predictive distribution of the $p(t(y^{\{(m)\}}) \mid y)$ compared to the value of the statistic evaluated on the observed data (for more details, see Gelman et al. (1996)). The advantage of posterior predictive checks is in their flexibility: they can be tailored to consider any desired aspect of a distribution.

Each of the models was implemented directly in the OpenBUGS software (Thomas et al., 2006). Chains were intialised by randomly drawing starting parameter values from their respective marginal priors. We consider a single, long–running chain for each model. We take a sequential confirmatory approach: initially a chain is updated for a small number of iterations (eg. 1000); autocorrelations and cross–correlations are assessed using the functionality provided in OpenBUGS and CODA (Plummer et al., 2008), and an appropriate thinning lag is decided upon; the Raftery and Lewis diagnostic (Raftery and Lewis., 1992) is employed to obtain an estimate of the burn-in period required in order to be able to estimate a predetermined quantile q to an accuracy of $\pm r$ with a certainty of s; the chain is updated according to the Raftery and Lewis suggested burn-in and the Heidelberger and Welch (Heidelberger and Welch., 1983) approach to convergence assessment is invoked and a decision is made as to how to proceed based on whether this diagnostic confirms convergence. For all the models in this study, we use $q = 0.025$, $r = \pm 0.0005$ and $s = 0.99$, so that in principle, we should be able to determine a two-sided, 95% CrI for the marginal posterior of each parameter to within 3 decimal places with 99% certainty. Once we are satisfied that convergence has been reached, we save the state of the chain and update the model for a further 1 million iterations, so that for example, an event that occurs as rarely as 1% of the time is expected to be simulated about ten–thousand times. This chain is used, via R, to produce posterior summaries of quantities of interest. Also, many replicate datasets are

simulated from the posterior predictive densities and these are used to assess model fit, both to the whole of the observed data and to the tail regions.

3. RESULTS

Table 2 presents the modes of the posterior predictive checks for the 10th, 50th (Median) and 90th percentiles. A 95% CrI is included for each of the model parameters, along with the DIC for each model.

In terms of overall fit, the gamma model shows the smallest DIC, followed closely by the Weibull model. The log–normal model shows the next best overall fit, and the exponential model shows the worst overall fit.

The PPCs reveal that the exponential model will over–predict the minimum and maximum, with probabilities 0.6 and 0.88 respectively, while it over–predicts the 10th, 50th and 90th percentiles with respective probabilities: 0.31, 0.54 and 0.64. The gamma model over–predicts the minimum and maximum with respective probabilities: 0.77 and 0.87. It over–predicts the 10th, 50th and 90th percentiles with probabilities: 0.59, 0.55 and 0.55 respectively. The log–normal model shows a tendency to predict the minimum and maximum greater than the observed minimum and maximum with probabilities: 0.8 and 0.98 respectively. It predicts higher than the observed values for the 10th, 50th and 90th percentiles with respective probabilities: 0.56, 0.52 and 0.71. The Weibull model over–predicts the minimum, 10th, 50th and 90th percentiles, and the maximum with probabilities: 0.75, 0.51, 0.57, 0.52 and 0.85 respectively.

In the lower tail the log-normal model fits best (predicting 31 flight-hours (fhr) where the observed data is 35 fhr. In the region of highest density, the data median time between failures is 107 fhr, and this is followed most closely by the gamma model (also 107 fhr) and then the Weibull model (109 fhr). In the upper percentiles, we observe a data 90th percentile of 290 fhr, which is modeled best by the log–normal (293 fhr) and followed by the exponential model (300 fhr).

Table **1**. Bayesian model summaries: posterior modes of distributions of posterior predictive 10^{th}, 50^{th} and 90^{th} percentiles; 95% CrI estimates for model parameters; model DICs.

Model	PPC Percentile Modes			95% CrI	DIC
	10%	50%	90%		
M1	14	89	300	$\lambda \in (0.0061, 0.0097)$	870.4
M2	29	107	271	$\alpha \in (1.31, 2.36)$ $\lambda \in (0.0096, 0.019)$	861.0
M3	31	95	293	$\mu \in (4.35, 4.75)$ $\tau \in (0.9371, 1.801)$	865.4
M4	27	109	266	$\alpha \in (1.13, 1.61)$ $\lambda' \in (0.0003, 0.004)$	861.8
Data	**35**	**107**	**290**		

4. DISCUSSION

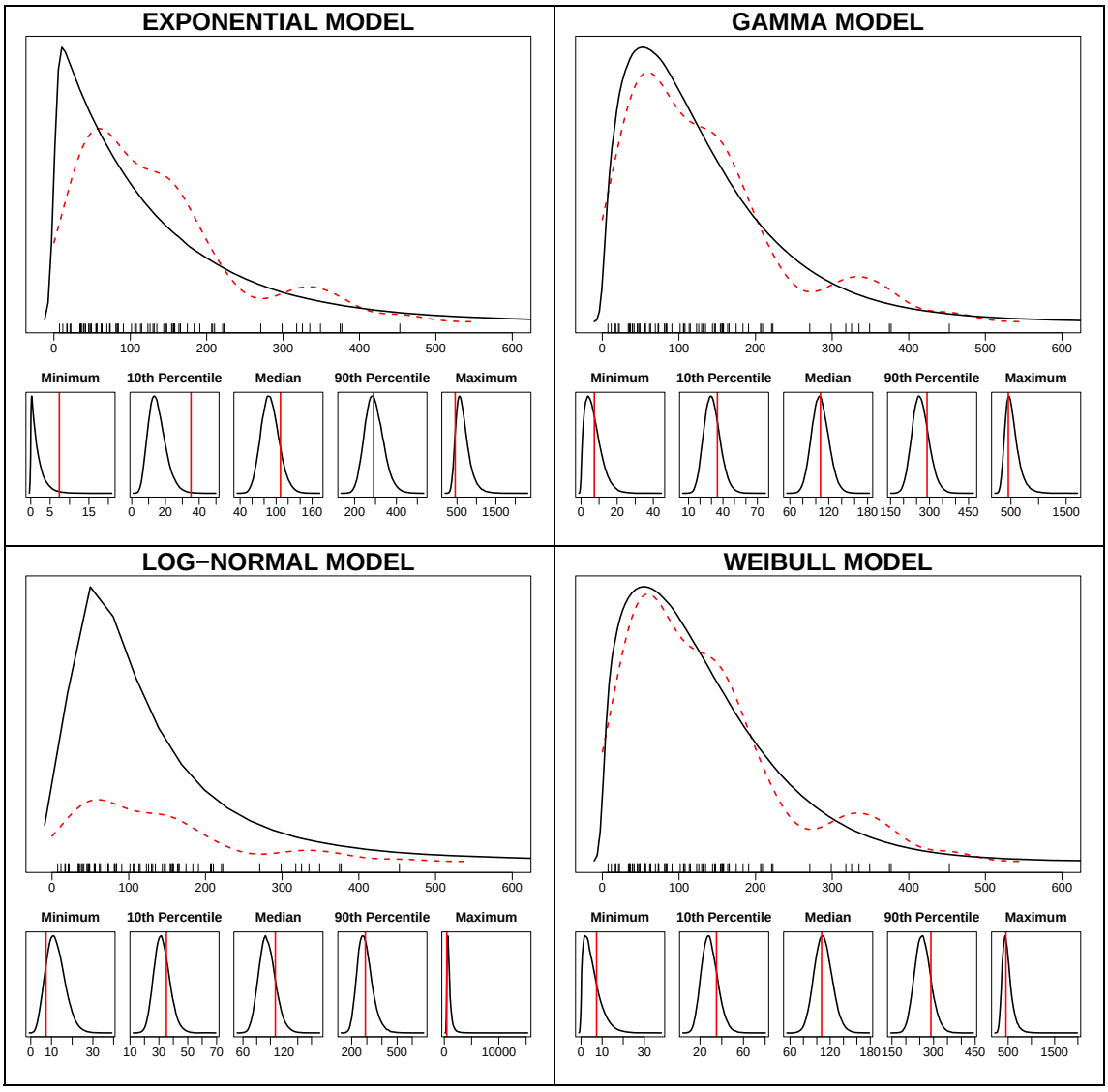
This paper asserts the need for specificity in model selection and comparison techniques so that models are chosen according to their suitability to the particular inferential task at hand. We build a case for such an assertion on the basis of a real example from the field of systems reliability.

We fit four models: exponential, gamma, log–normal and Weibull, to data representing inter–arrival times of failure events. The exponential model is a single parameter model while the other three models are 2 parameter models. Modeling is undertaken in a Bayesian setting where diffuse priors are employed so as to maximize the impact of the data models on the inferences produced. We demonstrate a combined approach to model selection which incorporates the use of the usual kind of penalized information measure (here we use DIC) as well as a suite of posterior predictive checks spanning the expected support of the data distribution.

We find that if the DIC had been our sole model selection criterion, the gamma model would be the chosen 'best' model. This is followed by the Weibull model, then the log–normal. The exponential model performs most poorly overall, but this is expected as it has only one parameter and therefore lacks a degree of parametric flexibility to fit data in relation to the two-parameter models. It should be noted that the conclusions listed here depend critically on the particular application and data used, although it is expected that they might also have wider implications.

When the aim of inference is concentrated on a particular region of the data support, we find that lower percentiles are best modeled by the log–normal distribution, followed by the gamma distribution. These same regions are described relatively poorly by the exponential and Weibull models. High–density (low to mid-percentile) regions are best modeled by the gamma distribution, although the Weibull model may also suffice in these regions. The log–normal model best describes the upper-percentile regions, followed by the exponential model. These percentiles are actually very poorly described by the gamma and Weibull models in comparison to the log-normal and exponential. This demonstrates the value in tailoring model selection to the inferential task at hand: we are better off choosing different models to describe distinct regions of interest.

Figure 1. Kernel density estimates of the observed data (dashed lines) and Posterior Predictive densities (solid lines) for the exponential, gamma, log-normal and Weibull models fitted to aircraft air-conditioner data. Plots of PPC densities are included for each model, for the predicted minimum, 10th percentile, median, 90th percentile and maximum values; vertical lines represent the corresponding values from the observed data.



The Bayesian approach to modeling via MCMC methods has been demonstrated to be a coherent, easily employed and easily manipulated framework for model fitting and model selection. The benefits of this approach are many and varied. They include its ease of implementation and the coherent interpretation of results as probability distributions and the ability to include prior information from other sources when it is available. While other methods of analysis make use of 'plug–in' estimates of parameters to assess tail fit, in practice full information is not available for the model parameters and hence the need for a model in the first place. Our approach incorporates full uncertainty about the true value of the parameters by producing probabilistic inferences given observed data. This is a far more open, straightforward approach to model performance assessment.

ACKNOWLEDGMENTS

The authors are grateful to adjunct Prof. Aitkin for the data. They also acknowledge the Centre for Collaborative Research on National Plant Biosecurity, the Institute for Sustainable Research and ARC Discovery Grant DP0667168 who supported this study.

REFERENCES

Alston, C., Kuhnert, P., Low Choy, S., McVinish, R. and Mengersen, K. (2005) Bayesian Model Comparison: Review and Discussion, *In Proceedings*: International Statistical Conference (ISI) 2005.

Al-Garni, A.Z., Tozan, M., Al-Garni, A.M., and Jamal, A., (2006). Failure Forecasting of Aircraft Air-Conditioning / Cooling Packs with Field Data. *DOI:*10.2514/1.26561.

Berry, D.A., (1996), Statistics: A Bayesian Perspective, *Duxbury Press Belmont, CA.*

Bolstad, W. M. (2007). Introduction to Bayesian Statistics. *Wiley-Interscience.*

Carlin, B. and Chib, S. (1995) Bayesian model choice via Markov chain Monte Carlo methods. *J. Royal Statist. Society Series B*, 57(3), 473-484.

Congdon, P., (2006). Bayesian Statistical Modelling, *John Wiley and Sons*, West Sussex.

Plummer, M., Best, N., Cowles, K., and Vines, K., (2008). CODA: Output Analysis and Diagnostics for MCMC. *R package version 0.133.*

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B., (1996a). Bayesian Data Analysis (2nd ed). *Chapman and Hall/CRC Press*, Florida.

Gelman, A., Meng, X.L., and Stern, H.S., (1996b). Posterior predictive assessment of model fitness via realized discrepancies with discussion. *Statistica Sinica* **6**, 733–807.

Geweke, J. (2005). Contemporary Bayesian Econometrics and Statistics, *Wiley Series in Probability and Statistics* (1st ed.). Wiley-Interscience.

Ghosh, J.K. (2006). An Introduction to Bayesian Analysis, *Springer-Verlag, New York.*

Gilks, W. R. (1995). Markov Chain Monte Carlo in Practice. *Chapman & Hall/CRC.*

Heidelberger, P., and Welch, P.D., (1983). Simulation Run Length Control in the Presence of an Initial Transient. *Operations Research*, 31(6), 1109-1144, URL *http://www.jstor.org/stable/170841*

Lee, K., Mengersen, K.L., Marin, J.-M., and Robert, C.P., (2008). Bayesian Inference on Mixtures of Distributions. *Proceedings of the Platinum Jubilee of the Indian Statistical Institute* (to appear).

Lee, P.M. (2004). Bayesian Statistics: An Introduction. (3rd ed.) Wiley, *ISBN: 978-0-340-81405-5*

Marin, J.-M., and Robert, C.P., (2007). Bayesian Core: A Practical Approach to Computational Bayesian Statistics. *Springer*, New York.

Thomas, A., O'Hara, B., Ligges, U., and Sturtz, S., (2006). Making BUGS Open. *R News* 6(1), 12-17.

Press, J.S., (2003). Subjective and Objective Bayesian Statistics: Principles, Models and Applications. *John Wiley and Sons*, New Jersey.

R Development Core Team., (2008). R:A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. ISBN: 3-900051-07-0, URL *http://www.R-project.org*.

Raftery, A.E., and Lewis, S., (1992). How many iterations in the Gibbs Sampler? In *Bayesian Statistics 4* (eds. J.M. Bernardo, J. Berger, A.P. Dawid and A.F.M. Smith), *Oxford: Oxford University Press*, 763-773.

Robert, C.P. (2007). The Bayesian Choice, *Springer-Verlag, New York.*

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and Van der Linde, A., (2002). Bayesian Measures of Model Complexity and Fit (with Discussion). *Journal of the Royal Statistical Society*, Series B, 64(4), 583-616.

Spiegelhalter, D.J., Abrams, K.R. and Myles, J.P. (2004). Bayesian Approaches to Clinical Trials and Health-Care Evaluation (3rd ed.), *John Wiley and Sons.*