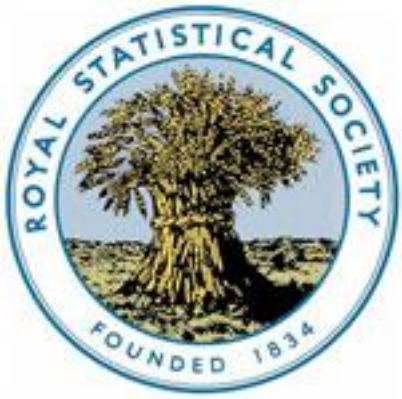



Bayesian Inference for Generalized Linear and Proportional Hazards Models via Gibbs Sampling

Author(s): P. Dellaportas and A. F. M. Smith

Source: *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, Vol. 42, No. 3 (1993), pp. 443-459
Published by: Wiley for the Royal Statistical Society
Stable URL: http://www.jstor.org/stable/2986324
Accessed: 20-10-2016 23:20 UTC



*Appl. Statist.* (1993)
**42**, *No.* 3, *pp.* 443–459

# Bayesian Inference for Generalized Linear and Proportional Hazards Models via Gibbs Sampling

By P. DELLAPORTAS

*University of Nottingham, UK*

and A. F. M. SMITH†

*Imperial College of Science, Technology and Medicine, London, UK*

[Received December 1990. Revised January 1992]

SUMMARY

It is shown that Gibbs sampling, making systematic use of an adaptive rejection algorithm proposed by Gilks and Wild, provides a straightforward computational procedure for Bayesian inferences in a wide class of generalized linear and proportional hazards models.

*Keywords*: Adaptive rejection algorithm; Bayesian inference; Generalized linear models; Gibbs sampling; Proportional hazards models; Quadratic logistic regression

## 1. Introduction

### 1.1. *Background*

A large statistical literature is devoted to generalized linear models (GLMs), as introduced by Nelder and Wedderburn (1972) and described in considerable detail in McCullagh and Nelder (1989). The same is true for proportional hazards models. It is striking, however, that very little of this literature involves Bayesian methodology. Leaving aside philosophical issues—and debates about whether GLMs should be interpreted as genuine models, or simply as exploratory data analysis devices—a major impediment to the routine Bayesian implementation of this large class of useful models has certainly been the difficulty of evaluating the integrals required.

Possible solutions include the numerical integration techniques introduced by Naylor and Smith (1982) (see Racine *et al.* (1986) and Grieve (1987, 1988)) or the analytical approximations suggested by Tierney and Kadane (1986) (see Albert (1988)). However, it seems that the perceived numerical or analytical sophistication involved in these solutions has inhibited their routine implementation. We illustrate the issues involved by discussing, in the next subsections, two concrete examples.

Our purpose in this paper is to demonstrate how an iterated simulation procedure—the Gibbs sampler—provides a relatively straightforward means of

†*Address for correspondence*: Department of Mathematics, Imperial College of Science, Technology and Medicine, Huxley Building, 180 Queen's Gate, London, SW7 2BZ, UK.

 0035–9254/93/42443 $2.00

making Bayesian inferences in a wide class of GLMs and proportional hazards models. The Gibbs sampler is implicit in the work of Hastings (1970), and made popular in the image processing context by Geman and Geman (1984). Detailed investigation of the applicability of the Gibbs sampling approach to general Bayesian calculation is given by Gelfand and Smith (1990), Gelfand *et al.* (1990), Carlin *et al.* (1992) and Gelfand *et al.* (1992).

### 1.2. *Quadratic Logistic Model*

Knuiman and Speed (1988) address the problem of incorporating prior information into the analysis of certain kinds of contingency table. They criticize the (unrealistic) approach of adopting a conjugate Dirichlet prior for reasons of analytic tractability and suggest instead the use of a non-conjugate multivariate normal prior distribution.

Table 1 consists of data, taken from Knuiman and Speed (1988), which reflect the relationship between duration of diabetes and retinopathy, an eye disease. Data from a previously reported study are also available in Table 1, serving as prior information for the parameters in the model. Knuiman and Speed suggest the use of a binomial GLM with a logit link function and a quadratic logistic model given by

$$\log\left(\frac{\pi_{1j}}{\pi_{2j}}\right) = \beta_1 + \beta_2 Z_j + \beta_3 Z_j^2 = \eta_j \qquad (1)$$

where $(\pi_{1j}, \pi_{2j})$ are the probabilities of being with or without retinopathy for persons with duration in the $j$th category, and $\mathbf{Z} = (Z_j)$ is the vector of middurations for the specified ranges, given by (1, 4, 7, 10, 13, 16, 19, 24).

They incorporate the prior information through a multivariate normal distribution, with mean vector and covariance matrix obtained from maximum likelihood estimates based on the prior data, and given by

$$\boldsymbol{\beta}_0 = \begin{pmatrix} -3.17 \\ 0.33 \\ -0.007 \end{pmatrix}, \qquad \mathbf{D}_0 = 10^{-4}\begin{pmatrix} 638 & & \\ -111 & 24.1 & \\ 3.9 & -0.9 & 0.04 \end{pmatrix}.$$

Model (1), combined with the prior $N(\boldsymbol{\beta}_0, \mathbf{D}_0)$, yields the joint posterior density

TABLE 1
*Diabetic retinopathy data from Knuiman and Speed (1988)*

| *Duration of diabetes (years)* | *Prior data: retinopathy* | | *Current data: retinopathy* | |
|---|---|---|---|---|
| | *Yes* | *No* | *Yes* | *No* |
| 0–2 | 17 | 215 | 46 | 290 |
| 3–5 | 26 | 218 | 52 | 211 |
| 6–8 | 39 | 137 | 44 | 134 |
| 9–11 | 27 | 62 | 54 | 91 |
| 12–14 | 35 | 36 | 38 | 53 |
| 15–17 | 37 | 16 | 39 | 42 |
| 18–20 | 26 | 13 | 23 | 23 |
| 21+ | 23 | 15 | 52 | 32 |

$$p(\boldsymbol{\beta}|\text{data}) \propto \exp\left[ -\tfrac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\mathbf{D}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) \sum_{j=1}^{8} \{x_{1j}\log\eta_j - (x_{1j}+x_{2j})\log(1+\exp\eta_j)\}\right] \quad (2)$$

where $x_{1j}$ and $x_{2j}$, $j=1, \ldots, 8$, are the current data numbers in each age category with and without retinopathy.

An examination of the posterior form (2) clearly reveals the implicit need for three-dimensional numerical integration to obtain the normalizing constant, together with repeated two-dimensional numerical integrations to summarize univariate marginals for $\beta_1$, $\beta_2$ and $\beta_3$.

Knuiman and Speed (1988) instead opted for an approximation, using a normal approximation based on the posterior mode, a solution of

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\beta}}\log p(\boldsymbol{\beta}|\text{data}) = 0,$$

and a measure of dispersion given by the matrix

$$D(\boldsymbol{\beta}) = -\left[\frac{\mathrm{d}^2\{\log p(\boldsymbol{\beta}|\text{data})\}}{\mathrm{d}\boldsymbol{\beta}\,\mathrm{d}\boldsymbol{\beta}^{\mathrm{T}}}\right]^{-1}$$

evaluated at the posterior mode. We shall demonstrate in Section 4 how a fully Bayesian numerical analysis of such a model can be given by using the Gibbs sampler, thus avoiding both direct numerical integration and the act of faith implicit in the assumption of approximate posterior normality.

### 1.3. *Weibull, Exponential and Extreme Value Proportional Hazards Models*

The Weibull, exponential and extreme value proportional hazards models are widely used for modelling censored survival data in which the response variate is the lifetime of a component or the survival time of a patient; see Aitkin and Clayton (1980). They represent important cases of a wider family of models, the proportional hazards models, introduced by Cox (1972). These models specify that the hazard function $\lambda(t; \mathbf{Z})$, for an individual with covariates $\mathbf{Z}$ at time $t$, is given by

$$\lambda(t; \mathbf{Z}) = \lambda_0(t)\exp(\mathbf{Z}\boldsymbol{\beta}), \quad (3)$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ is a vector of unknown parameters and $\lambda_0(t)$ is an unknown function of time. If $\lambda_0(t) = \lambda$, equation (3) reduces to the exponential regression model; with $\lambda_0(t) = \rho t^{\rho-1}$, we obtain the Weibull regression model, where $\rho$ is the shape parameter ($\rho > 0$); with $\lambda_0(t) = \alpha\exp(\alpha t)$ we obtain the extreme value regression model. Both the density and the survivor functions are needed for the development of the likelihood function based on uncensored and censored data. Denoting by $\Lambda(t)$ the cumulative hazard function, the survivor function is given by

$$S(t) = 1 - F(t) = \exp\{-\Lambda(t)\exp(\mathbf{Z}\boldsymbol{\beta})\}$$

where $F(t)$ denotes the cumulative density function, and the density function is given by

$$f(t) = \lambda_0(t)\exp\{\mathbf{Z}\boldsymbol{\beta} - \Lambda(t)\exp(\mathbf{Z}\boldsymbol{\beta})\}.$$

Consequently, the likelihood under the Weibull model is given by

$$\exp L(\boldsymbol{\beta}, \rho \mid \text{data}) = \left\{ \prod_{j=1}^{n} \rho t_j^{\rho-1} \exp(\mathbf{Z}_j \boldsymbol{\beta}) \right\} \left\{ \prod_{j=1}^{n+m} \exp\{ - t_j^{\rho} \exp(\mathbf{Z}_j \boldsymbol{\beta}) \right\} \quad (4)$$

where $t_j, j = 1, \ldots, n$, and $t_j, j = n+1, \ldots, m$, denote the uncensored and censored lifetimes respectively, $\rho$ is the shape parameter of the Weibull distribution ($\rho > 0$) and $\mathbf{Z}_j, j = 1, \ldots, m$, is the vector of covariates for the $j$th case.

The data in Table 2 are from Grieve (1987), who presented a Bayesian analysis using a Weibull regression model with proportional hazards, the calculations being carried out by using the adaptive quadrature techniques of Naylor and Smith (1982). The likelihood is given by equation (4), where

(a) $z_0 = 1$ for all mice,
(b) $z_1 = 1$ for mice in the vehicle control group and $z_1 = 0$ otherwise,
(c) $z_2 = 1$ for mice in the test substance group and $z_2 = 0$ otherwise, and
(d) $z_3 = 1$ for mice in the positive control group and $z_3 = 0$ otherwise,

so that the parameters of interest, $\beta_1$, $\beta_2$ and $\beta_3$, represent, respectively, the differential effects (from the irradiated control group) of groups II, III and IV described in Table 2, and following Grieve (1987) the $n+m = 80$ times in Table 2 are ordered in such a way that the first $n = 65$ times $t_1, t_2, \ldots, t_n$ are uncensored and the last $m = 15$ times, $t_{n+1}, t_{n+2}, \ldots, t_{n+m}$, are censored (i.e. correspond to deaths). For

TABLE 2
*Photocarcinogenicity data from Grieve (1987)*

| *Mouse* | *I, irradiated control* | | *II, vehicle control* | | *III, test substance* | | *IV, positive control* | |
|---|---|---|---|---|---|---|---|---|
| | *Week of death (censoring time)* | *Week of tumour* | *Week of death (censoring time)* | *Week of tumour* | *Week of death (censoring time)* | *Week of tumour* | *Week of death (censoring time)* | *Week of tumour* |
| 1 | | 12 | | 32 | | 22 | | 27 |
| 2 | | 17 | | 27 | | 26 | | 18 |
| 3 | | 21 | | 23 | 10 | | | 22 |
| 4 | | 25 | | 12 | | 28 | | 13 |
| 5 | | 11 | | 18 | | 19 | | 18 |
| 6 | | 26 | 40 | | | 15 | | 29 |
| 7 | | 27 | 40 | | | 12 | | 28 |
| 8 | | 30 | | 38 | | 35 | 20 | |
| 9 | | 13 | | 29 | | 35 | | 16 |
| 10 | | 12 | | 30 | | 10 | | 22 |
| 11 | | 21 | 40 | | | 22 | | 26 |
| 12 | | 20 | | 32 | | 18 | | 19 |
| 13 | | 23 | 40 | | 24 | | 29 | |
| 14 | | 25 | 40 | | | 12 | 10 | |
| 15 | | 23 | 40 | | 40 | | | 17 |
| 16 | | 29 | 40 | | 40 | | | 28 |
| 17 | | 35 | | 25 | | 31 | | 26 |
| 18 | 40 | | | 30 | | 24 | | 12 |
| 19 | | 31 | | 37 | | 37 | | 17 |
| 20 | | 36 | | 27 | | 29 | | 26 |

illustration, we shall assume the prior specification $p(\boldsymbol{\beta}, \rho) = \text{constant}$.

An examination of equation (4) clearly reveals the implicit need for five-dimensional numerical integration to obtain the normalizing constant, together with repeated four-dimensional numerical integrations to summarize univariate marginals for $\beta_0, \ldots, \beta_3, \rho$. Further challenging numerical tasks arise if, for example, we require univariate summary inferences for the survivor functions or median survival times in each of the groups. The adaptive quadrature approach of Naylor and Smith (1982) would then require a reparameterization for each such task, including the required function as a parameter and rederiving Jacobians. We shall demonstrate in Section 4 how the Gibbs sampler approach enables any required marginal inference summary to be obtained without the need for restarting the computational task from scratch.

### 1.4. *Outline of the Paper*

In Section 2, we give a brief account of the Gibbs sampling approach. In Section 3, we show that for a wide range of GLMs and proportional hazards models the joint posterior density is log-concave. This property can be exploited to obtain easily the required simulated samples for the application of the Gibbs sampling approach. In Section 4, we illustrate our procedures with the quadratic logistic and Weibull proportional hazards problems introduced in Sections 1.2 and 1.3.

## 2. Gibbs Sampler

### 2.1. *General Description*

The Gibbs sampler is a Markovian updating scheme enabling one to obtain (in the limit) samples from a joint distribution, via iterated sampling from full conditional distributions. Given a joint posterior density $p(\boldsymbol{\theta}|\mathbf{x})$, functional forms of the $k$ univariate full conditional densities (i.e. the distribution of each individual component of $\boldsymbol{\theta}$ conditional on specified values of the data $\mathbf{x}$ and all the other components) can be readily written down, at least up to proportionality. If, suppressing for notational convenience, the dependence on $\mathbf{x}$, these full conditional densities are denoted by

$$\left.\begin{array}{l} p(\theta_1|\theta_2, \theta_3, \ldots, \theta_k), \\ p(\theta_2|\theta_1, \theta_3, \ldots, \theta_k), \\ \quad\vdots \\ p(\theta_k|\theta_1, \theta_2, \ldots, \theta_{k-1}), \end{array}\right\} \tag{5}$$

then the Gibbs sampling algorithm proceeds as follows: choose initial values $\theta_2^{(0)}, \theta_3^{(0)}, \ldots, \theta_k^{(0)}$ and generate a value $\theta_1^{(1)}$ from the conditional density

$$p(\theta_1|\theta_2^{(0)}, \theta_3^{(0)}, \ldots, \theta_k^{(0)}).$$

Similarly, generate a value $\theta_2^{(1)}$ from the conditional density

$$p(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, \ldots, \theta_k^{(0)}).$$

and continue up to the value $\theta_k^{(1)}$ from the conditional density

$$p(\theta_k|\theta_1^{(1)}, \theta_2^{(1)}, \ldots, \theta_{k-1}^{(1)}).$$

Then, with the new realization of $\boldsymbol{\theta}$, $\boldsymbol{\theta}^{(1)}$, replacing the initial values the above process is iterated, say $t$ times, producing $\boldsymbol{\theta}^{(t)}$. Under mild conditions, as shown by Geman and Geman (1984) for discrete distributions, we have

$$\theta_i^{(t)} \xrightarrow{\mathrm{d}} \theta_i \sim p(\theta_i)$$

and therefore, for large $t$, $\theta_i^{(t)}$ can be regarded (approximately) as a simulated observation from $p(\theta_i)$, the marginal distribution of $\theta_i$. Independent parallel replication of this entire process $N$ times produces $N$ sets of parameter vectors ($\theta_j^{(t)} \equiv \theta_j, j = 1, \ldots, N$) and thus for each element of $\boldsymbol{\theta}$ we obtain a simulated sample of size $N$ from its marginal density. Although all computer-generated random variates are discrete (by virtue of hardware limits of arithmetic accuracy), we shall follow conventional practice in stochastic simulation and refer, without further qualification, to implicitly discretized versions of continuous distributions as 'continuous'. In such continuous cases, reconstruction of marginal densities can proceed by using either a kernel density estimate (see Silverman (1986)) or by averaging over the conditional density:

$$p(\theta_i) = \frac{1}{N} \sum_{j=1}^{N} p(\theta_i | \theta_j^{(t)}, j \neq i).$$

See Gelfand and Smith (1990).

Sophisticated numerical (see, for example, Smith *et al.* (1987)) or analytical (see, for example, Tierney and Kadane (1986)) approximation techniques already exist for Bayesian calculation. The iterative algorithm described here does not generally compete with these other methods in terms of efficiency. However, it provides a method which is simple to implement and does not require numerical or analytical sophistication on the part of the user. It has been demonstrated that the method can be used successfully in otherwise numerically (and analytically) intractable problems (see, for example, Coursaget *et al.* (1991), Carlin *et al.* (1992), Gelfand and Smith (1990) and Gelfand *et al.* (1990, 1992)).

The Gibbs sampler involves drawing random samples from all full conditional densities of the form (5). Often the likelihood and prior forms specified in Bayesian analysis lead to distributions in expression (5) which are of a familiar form, such as normals or gammas; see for example Gelfand and Smith (1990) and Gelfand *et al.* (1990). In these cases, standard algorithms are available to generate random variates; see, for example, Devroye (1986) or Ripley (1987).

In other cases, random variate generating methods such as the 'inversion method', the 'rejection method' or the 'ratio of uniforms method', applicable to wide ranges of distributions, can be used; see Devroye (1986), chapters 2–4 and 7, for a detailed description of such methods. However, depending on the nature of the distribution family, an efficient choice of a method generally requires mathematical insight on the part of the designer of the sampling scheme, e.g. exploiting a property such as log-concavity, or the knowledge of certain density characteristics such as the supremum of the density or the explicit form of the inverse of the cumulative density function. In addition, owing to their 'universality', these methods do not compete in efficiency with special purpose algorithms designed for the generation of random variates from popular densities. It is evident therefore that special care must be taken in both the choice and the design of such methods in the application of Gibbs sampling. For

GLMs and the proportional hazards models considered here, the key is provided by the following.

### 2.2. *Rejection Sampling from Log-concave Density Functions*

An important family of density functions which we shall consider in this subsection is the family of log-concave density functions. This family includes many common probability density functions: see, for example, Gilks and Wild (1992) or Devroye (1986), p. 287. We begin with a formal definition of what is meant by log-concavity. This is followed by a description of a specific rejection sampling method for dealing with log-concave density functions.

A positive function $f$ on an open convex set $C$ in $\mathbf{R}^n$ is called log-concave if $\log f$ is a twice continuously differentiable real-valued function on $C$, and its Hessian matrix

$$H = (H_{ij}(x)), \qquad H_{ij}(x) = \frac{\partial^2(\log f)}{\partial x_i \partial x_j}(x_1, \ldots, x_n)$$

is negative semidefinite for every $x \in C$. If the Hessian matrix is negative definite, the function $f$ is called strictly log-concave.

The log-concavity of a density function enables us to use specifically designed algorithms for the generation of random variates. However, these methods, in general, require knowledge of the position of the mode, thus necessitating a time-consuming maximization step. See Devroye (1986), pages 287–309, for a clear account of many such available methods for sampling from log-concave density functions.

Recently, Gilks and Wild (1992) have proposed a novel 'adaptive rejection sampling' method of sampling from any log-concave univariate probability density function, which has the important advantage of avoiding such maximization. Their suggested algorithm is based on the fact that any concave function can be bounded by piecewise linear upper and lower bounds (hulls), constructed by using tangents at, and chords between, evaluated points on the function over its domain. The detailed procedure, which we have found to be particularly convenient for Gibbs sampling, is as follows.

Assume that we need to generate random variates from the univariate probability density function $p(x) \propto \exp M(x)$, say. Suppose that $M(x)$ and $M'(x) = \partial M/\partial x$ have been evaluated at $k$ ordered points $x_1, x_2, \ldots, x_k$, let $T_k = [x_i, i = 1, \ldots, k]$ and denote the upper and lower hulls of $M(x)$ derived from the $k$ points by $u_k(x)$ and $l_k(x)$ respectively. Assume also that the mode of $M(x)$ lies between $x_1$ and $x_k$, that $M(x)$ is twice continuously differentiable on a real interval $(a, b)$, where $a$ and $b$ can be $-\infty$ or $\infty$, and that the second derivative is non-positive throughout $(a, b)$. Define

$$S_k(x) = \exp u_k(x) \Big/ \int \exp u_k(x)\, dx$$

and proceed according to the following algorithm.

*Repeat* until desired number of points have been sampled
  Sample $x$ from $S_k(x)$ and independently $u$ from uniform(0, 1)
  *If* $u \leqslant \exp\{l_k(x) - u_k(x)\}$ *then*
    Accept $x$
  *Else*
    *If* $u \leqslant \exp\{M(x) - u_k(x)\}$ *then*

Accept $x$
*Else*
Reject $x$
*End if*
Add $x$ to $T_k$, increment $k$, relabel the members of $T_k$
*End if*
*End Repeat*

The adaptive rejection sampling algorithm has two important advantages compared with other existing general purpose methods for generating independent observations from a probability density function.

First, unlike the other existing methods for generating from log-concave density functions (Devroye (1986), pages 287–309) and most well-known universal random variate generating methods, such as the rejection sampling or the ratio of uniforms methods, location of the mode is not required. Except for some well-known densities, locating the mode necessitates the use of numerical maximization routines, which require, on average for the kinds of density arising from GLMs and proportional hazards models, seven or eight function evaluations. We note that Gilks and Wild (1992) report an average of three function evaluations to obtain a sample of size 1 using the adaptive rejection sampling algorithm (and, indeed, in our experiments, as we report later, we have recorded an average sample size of less than 4).

Secondly, it is adaptive in the sense that the rejection probability is decreasing as more random variates are sampled from the envelope function because, with the addition of more points, the density function is closer to the upper and lower functions used to squeeze it.

## 3. Log-concavity

### 3.1. *Generalized Linear Models*

In this section we investigate the potential use of Gibbs sampling for making Bayesian inferences about the parameters in a GLM. We shall use the adaptive rejection sampling technique introduced by Gilks and Wild (1992) and so begin by discussing the log-concavity of the likelihood function. (Our emphasis here is therefore rather different from that of Zeger and Karim (1991), who also consider Gibbs sampling for GLMs but are more concerned with extensions to the random effects case than in optimizing the efficiency of rejection sampling.)

Let the data consist of an $n$-vector of responses $\mathbf{y}$, and an $n \times p$ known matrix of regressors $\mathbf{Z}$. The responses $\mathbf{y}$ are assumed to be a realization of a vector of random variables $\mathbf{Y}$ independently distributed with means $\boldsymbol{\mu}$. GLMs are then characterized by the following structure (see, for example, McCullagh and Nelder (1989)).

(a) The distribution of an individual response $y_i$ is assumed to belong to a natural exponential family

$$f(y_i|\theta) = \exp[\{\theta y_i - b(\theta)\}/a(\phi) + c(y_i, \phi)]$$

for some functions $a(\ )$, $b(\ )$ and $c(\ )$, for a so-called canonical parameter $\theta$ and known scale-related parameter $\phi$.

(b) The matrix $\mathbf{Z}$ influences $\mathbf{y}$ via a linear combination $\boldsymbol{\eta} = \mathbf{Z}\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a $p$-dimensional parameter vector and $\boldsymbol{\eta}$ is the so-called linear predictor vector.

(c) The linear predictor $\eta$ is related to the mean $\mu$ of $\mathbf{Y}$ by a link function $g$, such that $\eta_i = g(\mu_i)$, $i = 1, \ldots, n$. Of special importance are the canonical link functions, which occur when $\theta = \mathbf{Z}\beta$, for the canonical parameter $\theta$.

As is well known, the above family includes some very familiar models. For example, for the normal distribution and the canonical (in this case, identity) link function $g(\mu) = \mu$ we obtain the classical linear regression model. For the Poisson distribution, the canonical link function $g(\mu) = \log \mu$ gives rise to the log-linear Poisson model, which can be used, for example, in the analysis of multidimensional contingency tables. When the responses follow the binomial distribution with mean $\pi$, the link functions $g(\pi) = \text{logit}(\pi)$, $g(\pi) = \Phi^{-1}(\pi)$ and $g(\pi) = \log\{-\log(1-\pi)\}$ yield the logistic, probit and the complementary log–log models respectively.

Given the log-likelihood function $L$ of a GLM, maximum likelihood estimators are frequently used to estimate the vector $\beta$ of coefficients of the linear combination $Z\beta$; see, for example, McCullagh and Nelder (1989). Resulting test and confidence interval methods rely heavily on the asymptotic properties of the maximum likelihood estimators as the sample size $n$ of observations tends to infinity. Regularity conditions have been given by various researchers which guarantee, at least for canonical link functions, weak consistency and asymptotic normality of the maximum likelihood estimators; see, for example, Haberman (1977) and Fahrmeir and Kaufmann (1985). Typically, these regularity conditions ensure that the Fisher information matrix is positive definite. This is of great importance for the Bayesian approach that we follow here because, for canonical link functions, the Hessian and the information matrix coincide (see McCullagh and Nelder (1989), p. 43), and therefore log-concavity of the log-likelihood is automatically implied.

This result cannot be straightforwardly generalized for non-canonical link functions. However, Wedderburn (1976) provides a series of special cases in which he proves log-concavity of the likelihood function. His results are summarized as follows:

(a) *normal*—$L$ is strictly concave only for the canonical link function, where $g$ is the identity function;
(b) *gamma*—strict log-concavity is attained for $g(\mu) = \log \mu$ and $g(\mu) = \mu^\gamma$ $(-1 \leqslant \gamma < 0)$; it is assumed here that $y_i \geqslant 0$ for every $i$, $i = 1, 2, \ldots, n$);
(c) *Poisson*—$L$ is strictly concave for $g(\mu) = \log \mu$ and $g(\mu) = \mu^\gamma (0 < \gamma < 1)$; for the link function $g(\mu) = \mu$ the log-likelihood is strictly concave if $y_i > 0$ for every $i$, and concave for any value of $y_i$;
(d) *binomial*—the logistic, probit and complementary log-log models defined above attain strict log-concavity of the likelihood function; $L$ is also strictly concave for the link functions $g(\mu) = \mu$, and $g(\mu) = \sin^{-1}\sqrt{\mu}$.

An interesting point arises here. Wedderburn (1976) shows that for the logistic, probit and complementary log–log models the maximum likelihood estimators are guaranteed to be finite only when $0 < y_i < m_i$ for every $i$, where $y_i$ is the number of positive responses from $m_i$ trials. In addition, for the last two link functions, $g(\mu) = \mu$ and $g(\mu) = \sin^{-1}\sqrt{\mu}$, the finiteness of the maximum likelihood estimates is not guaranteed. However, in the Bayesian context, the prior distribution typically overcomes this problem, yielding a well-behaved posterior density. Consequently, the difficulties which arise in the maximum likelihood estimation approach do not usually occur when a Bayesian approach with a suitable prior is adopted.

The application of the adaptive rejection sampling method described in Section 2.2 requires the log-concavity of the full conditional distributions. We have summarized conditions and cases where the likelihood of a GLM is log-concave viewed as a function of the complete parameter vector. We remark that this statement implies log-concavity of the full conditional likelihoods (i.e. the likelihood considered as a function of a single component of $\theta$ for fixed values of the other components) because, by definition, a negative semidefinite Hessian matrix has non-positive diagonal elements.

### 3.2. *Proportional Hazards Models*

Recall the likelihood for the Weibull proportional hazards model given in Section 1.3. Then, with $\log \mu_j = \rho \log t_j + \mathbf{Z}_j\boldsymbol{\beta}$, simple manipulation yields, for all $\rho$ and $\boldsymbol{\beta}$,

$$\frac{\partial^2 L}{\partial \beta_k^2} = -\sum_j z_{kj}^2 \mu_j \leqslant 0, \qquad 0 \leqslant k \leqslant p, \tag{6}$$

$$\frac{\partial^2 L}{\partial \rho^2} = -\frac{n}{\rho^2} - \sum (\log t_j)^2 \mu_j < 0.$$

Thus condition (6) guarantees log-concavity under the Weibull model and hence for the exponential model, the special case of the Weibull model with $\rho = 1$. Finally, the transformation $u = \exp t$ in equation (4) yields the extreme value distribution, so log-concavity is readily shown for this model also.

In a Bayesian analysis, the full conditional posterior log-density function for a specific parameter has the form of the sum of the corresponding full conditional log-likelihood and the corresponding logarithm of the full conditional prior density function. Consequently, if the prior density function is log-concave, the full posterior conditional will be log-concave, since we have a sum of two log-concave functions; see, for example, Rockafellar (1970).

For the remainder of this paper, we shall assume that the prior density functions used are log-concave, so that the conditions for the application of the adaptive rejection sampling algorithm described in Section 2.2 are fulfilled for many—including all the most common—GLMs and proportional hazards models.

In cases where the prior is not log-concave, a sampling–resampling technique described in Smith and Gelfand (1992) can be adopted. We shall not consider this case further; see Smith and Gelfand (1992) for more details. Also, we shall not discuss extensions to the case of unknown $\phi$ in the GLM case.

A referee has suggested that readers familiar with the way in which Aitkin and Clayton (1980) maximize the likelihood function may derive further intuition about the functioning of the Gibbs sampler by noting an analogy. Aitkin and Clayton proceed by Gauss–Siedel iteration between two sets of maximum likelihood equations: those for $\boldsymbol{\beta}$ conditional on $\rho$ (which, effectively, reduces to a GLM problem) and that for $\rho$ conditional on $\boldsymbol{\beta}$. The Gibbs sampler exploits the same iteration structure, but in terms of stochastic simulation rather than maximization. It is worth also noting that the rate of convergence of both the Gauss–Siedel and the Gibbs sampling procedures can be markedly affected by the parameterization used. For example, the more orthogonal the working parameters, typically the faster the convergence is. See Hills and Smith (1992) for a more detailed discussion of parameterization issues in Gibbs sampling.

## 4. Illustrative Examples

We shall consider Bayesian reanalyses, using the Gibbs sampler approach, of the two problems introduced in Sections 1.2 and 1.3.

The first example involves the quadratic logistic model. This will be used to illustrate the following important aspects of the fully Bayesian analyses that we are presenting: first, that the incorporation of an informative prior specification leads to no complication of the computational implementation; secondly, that the computation proceeds in an identical manner for both large and small samples, and involves no act of faith (implicitly large sample) approximations.

The second example involves the Weibull proportional hazards model. This will be used to the following effect: first, to illustrate further the calculation via the Gibbs sampler of full posterior densities for the model parameters; secondly, to illustrate the ease with which the output from the Gibbs sampling procedure can be flexibly used to present inference summaries for functions of the model parameters (e.g. median survival time or the full survivor function for specified subgroups).

Since this paper is primarily about the implementation of the Gibbs sampler, our focus here is entirely on inference and not on issues such as model checking, residual plots etc., which would be part of the wider strategy of using these models.

### 4.1. *Reanalysis of Quadratic Logistic Problem*

Recall the form of the joint posterior density for the model parameters $\beta_1$, $\beta_2$ and $\beta_3$, given by density (2). By virtue of our earlier discussion, this joint posterior yields log-concave full conditional posterior densities for each of $\beta_1$, $\beta_2$ and $\beta_3$, and the Gibbs sampling procedure of Section 2.1 is straightforwardly implemented by using the adaptive rejection algorithm of Section 2.2.

The full conditionals are identified, up to proportionality, by simply taking the form of $p(\boldsymbol{\beta}|\text{data})$ and regarding it, successively, as functions of $\beta_1$, $\beta_2$ and $\beta_3$ for specified values of the other two parameters. To initialize the Gibbs sampling procedure, maximum likelihood estimates of $\beta_1$, $\beta_2$ and $\beta_3$ obtained from a GLIM (Payne, 1985) analysis of the current data were used, replications of the iterative procedure thereafter proceeding independently.

At each iteration of the Gibbs sampling algorithm, adaptive rejection sampling from every conditional density requires (at least) two points which can be used as initial points for the construction of upper (using tangents) and lower (using chords) bounds. These initial points were taken as the sample means plus or minus one standard deviation, where the sample moments were calculated from the previous iteration of Gibbs sampling. In cases where the two initial points did not lie on either side of the mode of the conditional density, additional points were supplied.

An assessment of convergence of the process was made by using the pragmatic checks on stationarity outlined in Gelfand *et al.* (1990). In particular, a number of summary statistics from the replicated samples (first and second moments and selected percentiles) were monitored every 10 iterations for each parameter, together with $Q$–$Q$ plots, augmented by direct graphical comparison of successive marginal density reconstructions in the final stages once stationarity appeared to be achieved. For the initial iterations, relatively small numbers of replications (15–25, say) can be used to identify when the early, highly transient, phase of sampling is starting to stabilize. The number of replications can then be increased by spawning several new

iteration tracks from some or all of the initial tracks. Again, there will be an obviously transient phase which will subsequently settle down. Towards the end of the process, one might aim at, say, 100–500 replications, depending on the accuracy of posterior estimation required, or the detail required with graphical reconstructions of univariate or bivariate densities. In this example, the sampler was run for 150 iterations, by which time convergence had clearly occurred, and final iterations used 500 replications to ensure appropriate accuracy in a numerical comparison of estimators with those of Knuiman and Speed (1988).

The resulting estimated posterior mean vector and covariance matrix from the Gibbs sampler analysis (i.e. the usual sample estimates from the 500 sampled triples of $\beta_1$, $\beta_2$ and $\beta_3$ at the 150th iteration) are given by

$$\boldsymbol{\beta}^* = \begin{pmatrix} -2.36 \\ 0.21 \\ -0.004 \end{pmatrix}, \qquad \mathbf{D}^* = 10^{-4}\begin{pmatrix} 201 & & \\ -35.7 & 7.9 & \\ 1.2 & -0.3 & 0.01 \end{pmatrix},$$

which are in very close agreement with the posterior mean and covariance estimates $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{D}})$ reported by Knuiman and Speed (1988):

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} -2.37 \\ 0.21 \\ -0.004 \end{pmatrix}, \qquad \hat{\mathbf{D}} = 10^{-4}\begin{pmatrix} 207 & & \\ -36 & 8.1 & \\ 1.2 & -0.3 & 0.01 \end{pmatrix}.$$

This analysis verifies that the implicit assumption of Knuiman and Speed of approximate posterior normality was reasonable—although their approach provides no internal validation of this.

However, in many applications of GLMs, sample sizes are small and substantial prior information is not available. In such cases, assumptions of asymptotic normality can be misleading as we illustrate by the following analysis.

Suppose that the above problem had been somewhat different: no prior experiment had taken place, and the data were those in Table 3, derived from the real (Table 1) data set by (approximately) dividing each cell number by 50—fudging a little in the seventh row to avoid the (0, 0) case. Such experiments with small sample sizes are very common for example in the pharmaceutical industry, both on cost and ethical grounds. Consider the same quadratic logistic model, now with a non-informative

TABLE 3
*Diabetic retinopathy data (artificially generated)*

| *Duration of diabetes (years)* | *Retinopathy* | |
|---|---|---|
| | *Yes* | *No* |
| 0–2 | 1 | 6 |
| 3–5 | 1 | 4 |
| 6–8 | 1 | 3 |
| 9–11 | 1 | 2 |
| 12–14 | 1 | 1 |
| 15–17 | 1 | 1 |
| 18–20 | 1 | 0 |
| 21+ | 1 | 1 |

prior distribution $p(\beta_1, \beta_2, \beta_3) =$ constant. In this case, the posterior summary statistics proposed by Knuiman and Speed (1988) reduce to the maximum likelihood estimates, given by

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} -2.17 \\ 0.21 \\ -0.004 \end{pmatrix}, \qquad \hat{\mathbf{D}} = 10^{-4} \begin{pmatrix} 12894 & & \\ -2203 & 518.5 & \\ 74.7 & -19.9 & 0.84 \end{pmatrix}.$$

Gibbs sampling converged after about 300 iterations. Again, the maximum likelihood estimates were used as initial starting points, and the adaptive rejection sampling was used throughout the iterative sampling procedure. The sample mean vector and covariance matrix based on the final iteration and 500 replications were

$$\boldsymbol{\beta}^* = \begin{pmatrix} -2.48 \\ 0.25 \\ -0.005 \end{pmatrix}, \qquad \mathbf{D}^* = 10^{-4} \begin{pmatrix} 15048 & & \\ -2554 & 610.4 & \\ 87.9 & -24.3 & 1.1 \end{pmatrix}.$$

A visual inspection of the posterior marginal densities provides more striking insight into the inadequacy of the normal approximation, and the differences between the approximate (maximum likelihood) and the actual posteriors are clearly illustrated in Fig. 1.

Experience of different experimental runs suggests that, averaged over iterations of the Gibbs sampler, the adaptive sampling algorithm required somewhat under four function evaluations as the basis for constructing hulls to deliver a variate value from each full conditional density. Comparisons with other (maximization-based) techniques for sampling from log-concave densities suggests that the adaptive rejection algorithm provides a saving in function evaluations of about 50%.

### 4.2. *Reanalysis of Proportional Hazards Problem*

Recall the form of the joint posterior for $(\boldsymbol{\beta}, \rho)$ resulting from equation (4), together with the prior specification $p(\boldsymbol{\beta}, \rho) =$ constant.

We note that the forms of

$$p(\rho \mid \boldsymbol{\beta}, \text{data}) \quad \text{and} \quad p(\beta_i \mid \rho, \beta_j, j \neq i, \text{data}), \qquad i = 0, \ldots, 3,$$

are straightforwardly identified from equation (4). Again, the Gibbs sampler was initialized by using maximum likelihood estimates of $\rho$, $\beta_0, \ldots, \beta_3$ (derivable in GLIM, as in Aitkin and Clayton (1980)). Using a strategy similar to that described in Section 4.1, convergence was achieved within 150 iterations and marginal densities for $\beta_1$, $\beta_2$, $\beta_3$ and $\rho$, reconstructed from 500 final stage replications, are shown in Figs 2 and 3. The latter clearly rules out any simplifying assumption of an exponential survival distribution ($\rho = 1$). The former shows, in particular, that groups I and III (the vehicle and positive control groups) have very different survival distributions; $\beta_1 \neq \beta_3$.

As an illustration of the ease with which the Gibbs sampler approach now enables us to follow up this observation, without the need for substantial new computation, we present the following further analyses. Suppose that we wish to look at survival in these two groups in terms of the survivor functions,

$$S_j(t) = \exp\{-t^{\rho} \exp(\mathbf{Z}_j \boldsymbol{\beta})\}, \qquad j = 1, 3,$$

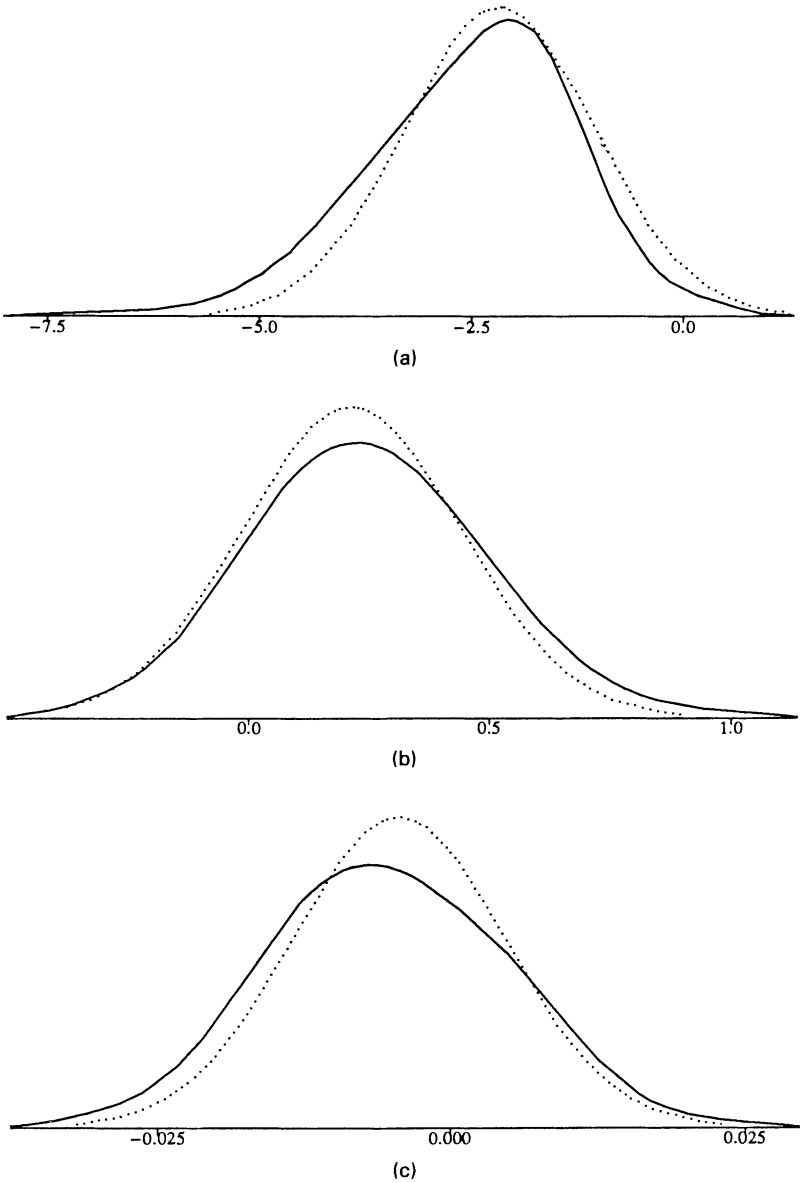


Fig. 1. Posterior marginals for (a) $\beta_0$, (b) $\beta_1$ and (c) $\beta_2$ (——, Gibbs sampler; ·····, maximum likelihood)

or to compare the median survival times ($m$, such that $S(m) = 0.5$),

$$m_j = \{\log 2 \exp(-\mathbf{Z}_j\boldsymbol{\beta})\}^{1/\rho}, \qquad j = 0, 1, 3,$$

for all three control groups.

Considering the latter first, we note that the random quantities $m_0$, $m_1$ and $m_3$ are simply specified non-linear functions of $\boldsymbol{\beta}$ and $\rho$. From the 500 final replicated sample

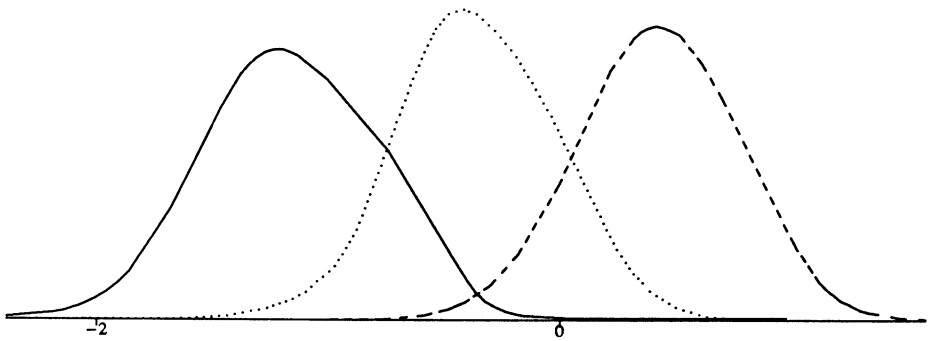


Fig. 2. Posterior marginals for $\beta_1$ (———), $\beta_2$ (·····) and $\beta_3$ (------)

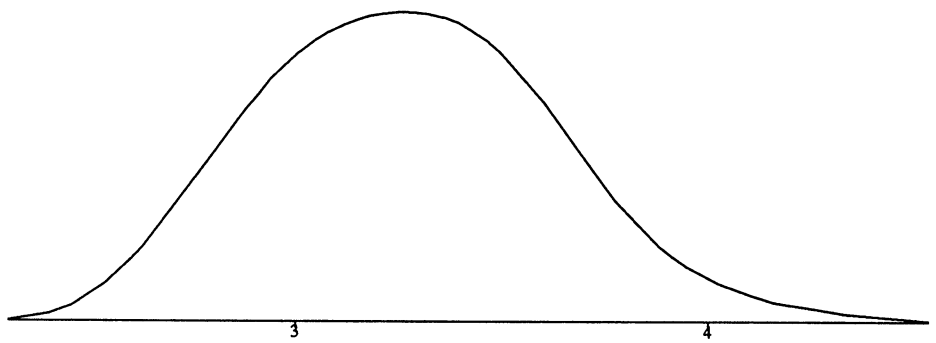


Fig. 3. Posterior marginal for $\rho$

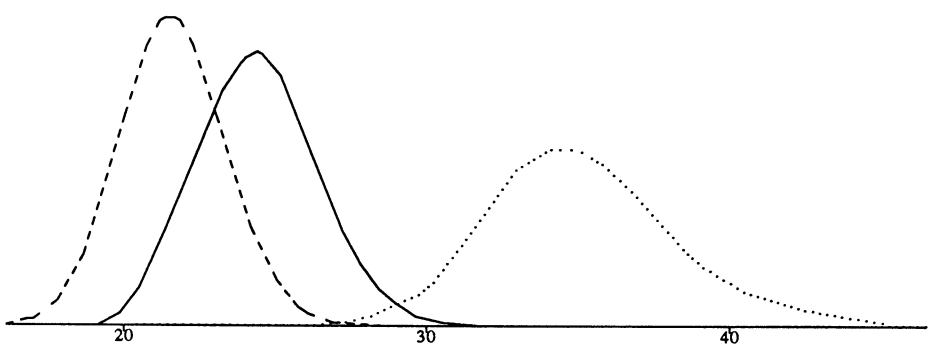


Fig. 4. Posterior median lifetimes: ———, irradiated control; ·····, vehicle control; ------, positive control

vectors of $(\beta_0, \beta_1, \beta_2, \beta_3, \rho)$, we can therefore immediately obtain 500 sampled values from each of the posteriors for $m_0$, $m_1$ and $m_3$. Fig. 4 displays the three marginal posteriors reconstructed from these samples. To give a summary visual description of posterior uncertainties about the two survivor functions for the vehicle and positive controls we proceed as follows. We choose a discrete range of five-weekly time values from 5 to 90 weeks. For each such specified time point $t$, $S_j(t)$, for $j = 1, 3$, is again just

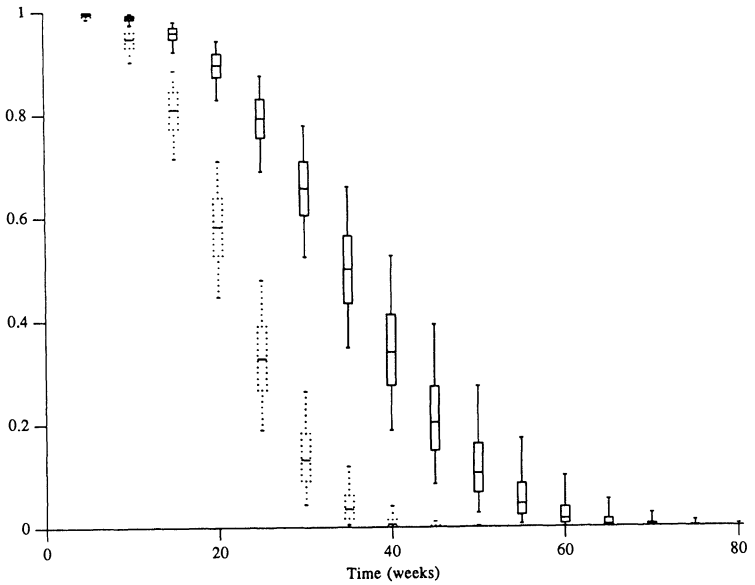


Fig. 5. Posterior survivor functions: ——, vehicle control; ·····, positive control

a non-linear function of $\beta$ and $\rho$. Again, therefore, directly from the final sample output of the Gibbs sampler we can generate a sample of 500 points from the posterior marginal for the point $S_j(t)$. These samples have been summarized in Fig. 5 in box plot form, indicating the 5th, 25th, 50th, 75th and 95th percentiles. The general form of posterior beliefs about the two survivor functions clearly emerges.

## Acknowledgements

The first author was supported under the Science and Engineering Research Council's Complex Stochastic Systems Initiative. Comments from the referees and Editor on an earlier version of the paper were extremely helpful.

## References

Aitkin, M. and Clayton, D. (1980) The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM. *Appl. Statist*, **29**, 156–163.

Albert, J. H. (1988) Bayesian estimation of Poisson means using a hierarchical log-linear model. In *Bayesian Statistics 3* (eds J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), pp. 519–531. New York: Oxford University Press.

Carlin, B. P., Gelfand, A. E. and Smith, A. F. M. (1992) Hierarchical Bayesian analysis of changepoint problems. *Appl. Statist.*, **42**, 389–405.

Coursaget, P., Yvonnet, B., Gilks, W. R., Wang, C. C., Day, N. E., Chiron, J.-P. and Diop-Mar, I. (1991) Scheduling of revaccination against hepatitis B virus. *Lancet*, **337**, 1180–1183.

Cox, D. R. (1972) Regression models and life-tables (with discussion). *J. R. Statist. Soc.* B, **34**, 187–220.

Devroye, L. (1986) *Non-uniform Random Variate Generation*. New York: Springer.

Fahrmeir, L. and Kaufmann, H. (1985) Consistency and asymptotic normality of the maximum likelihood estimator in Generalised Linear Models. *Ann. Statist.*, **13**, 342–368.

Gelfand, A. E., Hills, S. E., Racine-Poon, A. and Smith, A. F. M. (1990) Illustration of Bayesian inference in normal data models using Gibbs sampling. *J. Am. Statist. Ass.*, **85**, 972–985.

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling based approaches to calculating marginal densities. *J. Am. Statist. Ass.*, **85**, 398–409.

Gelfand, A. E., Smith, A. F. M. and Lee, T.-M. (1992) Bayesian analysis of constrained parameter and truncated data problems. *J. Am. Statist. Ass.*, **87**, 523–532.

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattn Anal. Mach. Intell.*, **6**, 721–741.

Gilks, W. R. and Wild, P. (1992) Adaptive rejection sampling for Gibbs sampling. *Appl. Statist.*, **41**, 337–348.

Grieve, A. P. (1987) Applications of Bayesian software: two examples. *Statistician*, **36**, 283–288.

——— (1988) A Bayesian approach to the analysis of LD50 experiments. In *Bayesian Statistics 3* (eds J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), pp. 617–630. New York: Oxford University Press.

Haberman, S. J. (1977) Maximum likelihood estimates in exponential response models. *Ann. Statist.*, **5**, 815–841.

Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Hills, S. E. and Smith, A. F. M. (1992) Parameterizations issues in Bayesian inference. In *Bayesian Statistics 4* (eds J. M. Bernardo, J. Berger, A. P. Dawid and A. F. M. Smith), pp. 227–246. Oxford: Oxford University Press.

Knuiman, M. W. and Speed, T. P. (1988) Incorporating prior information into the analysis of contingency tables. *Biometrics*, **44**, 1061–1071.

McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.

Naylor, J. C. and Smith, A. F. M. (1982) Applications of a method for the efficient computation of posterior distributions. *Appl. Statist.*, **31**, 214–225.

Nelder, J. A. and Wedderburn, R. W. M. (1972) Generalized linear models. *J. R. Statist. Soc.* A, **135**, 370–384.

Payne, C. D. (ed.) (1985) *The GLIM System: Release 3.77*. Oxford: Numerical Algorithms Group.

Racine, A., Grieve, A. P., Flühler, H. and Smith, A. F. M. (1986) Bayesian methods in practice: experiences in the pharmaceutical industry (with discussion). *Appl. Statist.*, **35**, 93–150.

Ripley, B. D. (1987) *Stochastic Simulation*. New York: Wiley.

Rockafellar, R. T. (1970) *Convex Analysis*, p. 77. Princeton: Princeton University Press.

Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Smith, A. F. M. and Gelfand, A. E. (1992) Bayesian Statistics without tears: a sampling–resampling perspective. *Am. Statistn*, **46**, 84–88.

Smith, A. F. M., Skene, A. M., Shaw, J. E. H. and Naylor, J. C. (1987) Progress with numerical and graphical methods for Bayesian statistics. *Statistician*, **36**, 75–82.

Tierney, L. and Kadane, J. (1986) Accurate approximations for posterior moments and marginals. *J. Am. Statist. Ass.*, **81**, 82–86.

Wedderburn, R. W. M. (1976) On the existence and uniqueness of the maximum likelihood estimates for certain Generalised Linear Models. *Biometrika*, **63**, 27–32.

Zeger, S. L. and Karim, M. R. (1991) Generalized linear models with random effects: a Gibbs sampling approach. *J. Am. Statist. Ass.*, **86**, 79–86.