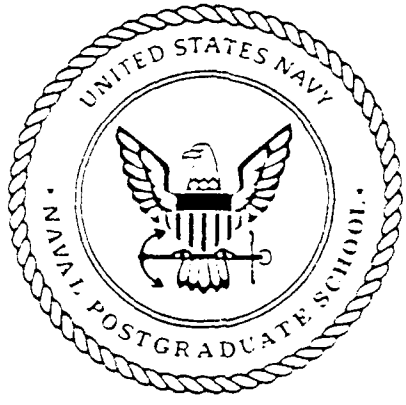
NPSOR-91-020

# NAVAL POSTGRADUATE SCHOOL

## Monterey, California






# BAYESIAN COMPUTATIONS IN SURVIVAL MODELS VIA THE GIBBS SAMPLER

Lynn Kuo and Adrian F. M. Smith

July 1991



Prepared for:
Naval Postgraduate School
Monterey, CA



**NAVAL POSTGRADUATE SCHOOL,**
**MONTEREY, CALIFORNIA**

Rear Admiral R. W. West, Jr.
Superintendent

Harrison Shull
Provost

This report was prepared for and funded by the Naval Postgraduate School.

This report was prepared by:

LYNN KUO
Professor of Operations Research

Reviewed by:

PETER PURDUE
Professor and Chairman
Department of Operations Research

Released by:

PAUL J. MARTO
Dean of Research

Unclassified

Security Classification of this page

# REPORT DOCUMENTATION PAGE

| | |
|---|---|
| 1a Report Security Classification UNCLASSIFIED | 1b Restrictive Markings |
| 2a Security Classification Authority | 3 Distribution Availability of Report: Approved for public release; distribution is unlimited |
| 2b Declassification/Downgrading Schedule | |
| 4 Performing Organization Report Number(s) NPSOR-91-20 | 5 Monitoring Organization Report Number(s) |
| 6a Name of Performing Organization: Naval Postgraduate School; 6b Office Symbol (*If Applicable*) OR | 7a Name of Monitoring Organization |
| 6c Address (*city, state, and ZIP code*): Monterey, CA 93943-5000 | 7b Address (*city, state, and ZIP code*) |
| 8a Name of Funding/Sponsoring Organization: Naval Postgraduate School; 8b Office Symbol (*If Applicable*) OR/Ku | 9 Procurement Instrument Identification Number: O&MN, Direct Funding |
| 8c Address (*city, state, and ZIP code*) Monterey, California | 10 Source of Funding Numbers |

| Program Element Number | Project No | Task No | Work Unit Accession No |
|---|---|---|---|
| | | | |
| | | | |

11 Title (*Include Security Classification*) Bayesian Computations in Survival Models via the Gibbs Sampler

12 Personal Author(s) Lynn Kuo and Adrian F. M. Smith

| 13a Type of Report | 13b Time Covered | 14 Date of Report (*year, month,day*) | 15 Page Count |
|---|---|---|---|
| Technical | From To | 1991, July | |

16 Supplementary Notation The views expressed in this paper are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

| 17 Cosati Codes | | |
|---|---|---|
| Field | Group | Subgroup |
| | | |
| | | |

18 Subject Terms (*continue on reverse if necessary and identify by block number*)
Censored data; truncated data; parametric survival model; nonparametric survival model; Bayesian computation

19 Abstract (*continue on reverse if necessary and identify by block number*

Survival models used in biomedical and reliability contexts typically involve data censoring, and may also involve constraints in the form of ordered parameters. In addition, inferential interest often focuses on non-linear functions of natural model parameters. From a Bayesian statistical analysis perspective, these features combine to create difficult computational problems by seeming to require (multi-dimensional) numerical integrals over awkwardly defined regions. This paper illustrates how these apparent difficulties can be overcome, in both parametric and non-parametric settings, by the Gibbs sampler approach to Bayesian computation.

| 20 Distribution/Availability of Abstract | 21 Abstract Security Classification |
|---|---|
| [X] unclassified/unlimited [ ] same as report [ ] DTIC users | Unclassified |

| 22a Name of Responsible Individual | 22b Telephone (*Include Area code*) | 22c Office Symbol |
|---|---|---|
| Lynn Kuo | (408) 646-2119 | OR/Ku |

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted

All other editions are obsolete

security classification of this page

Unclassified

[FILE: ASBCSM.CHI (JUNE 1991)]

# BAYESIAN COMPUTATIONS IN SURVIVAL MODELS VIA THE GIBBS SAMPLER

LYNN KUO,
Department of Statistics,
University of Connecticut,
U. S. A.

ADRIAN F. M. SMITH,
Department of Mathematics,
Imperial College London,
United Kingdom.

ABSTRACT. Survival models used in biomedical and reliability contexts typically involve data censoring, and may also involve constraints in the form of ordered parameters. In addition, inferential interest often focuses on non-linear functions of natural model parameters. From a Bayesian statistical analysis perspective, these features combine to create difficult computational problems by seeming to require (multi-dimensional) numerical integrals over awkwardly defined regions. This paper illustrates how these apparent difficulties can be overcome, in both parametric and non-parametric settings, by the Gibbs sampler approach to Bayesian computation.

## 1. Introduction

For the Bayesian statistical analysis of other than simple stylized models, the key tool for calculation is (multi-dimensional) numerical integration; see, for example, Smith *et al* (1987) for a review of available techniques. However, it is widely recognized that considerable numerical sophistication is typically required in applying these techniques, and that this has thus far hampered the development of routinely available, user-friendly, Bayesian computational methods.

This is particularly true in the case of survival models used in biomedical and reliability contexts. Here, features such as data censoring, ordered parameters, assumed convexity or concavity of distributions, all conspire to produce complicatedly constrained regions over which numerical integrations are required. Not surprisingly, the literature therefore contains very few instances of fully Bayesian analyses in survival contexts (*i.e.*, presenting full and accurate posterior summaries, rather than, say modal point estimates or second derivative uncertainty measures).

Recently, however, Gelfand *et al* (1991) have shown that the Gibbs sampler approach to Bayesian computation (see, for example, Gelfand and Smith, 1990, and Gelfand *et al*, 1990) effectively side-steps the

seeming problems of awkwardly defined integration regions in truncated data and constrained parameter problems, and provides an easily implemented computational procedure.

Our purpose in this paper is to illustrate the simplicity and scope of the Gibbs sampler for the routine Bayesian analysis of survival data, in both parametric and non-parametric settings. In Section 2, we briefly review the Gibbs sampler and its general structure for constrained parameter and censored data problems. In Section 3, we provide a range of illustrations of how the methodology proceeds for a variety of parametric models used in various survival modelling contexts. In Section 4, we give a non-parametric illustration of the methodology.

## 2. The Gibbs Sampler, Constraints and Censoring

### 2.1 THE GIBBS SAMPLER

In what follows, densities will be denoted generically by square brackets, so that joint, conditional and marginal forms for random variables $U$, $V$ appear, respectively, as $[U, V]$, $[U|V]$ and $[V]$. Marginalization by integration is denoted by $[U] = \int [U|V] \cdot [V]$. Given a collection of random variables with joint density $[U_1, U_2, \ldots, U_k]$, we shall refer to $[U_s|U_r, r \neq s]$, $s = 1, 2, \ldots, k$, as the full conditional densities.

The Gibbs sampler is a simply described iterative stochastic simulation scheme, whereby samples drawn from the full conditional densities are used to provide summaries of aspects of the joint density. Given an arbitrary set of starting values, $U_1^0, \ldots, U_k^0$, a random variate $U_1^1$ is drawn from $[U_1|U_2^0, \ldots, U_k^0]$, then a variate $U_2^1$ is drawn from $[U_2|U_1^1, U_3^0, \ldots, U_k^0]$, and so on until $U_k^1$ is drawn from $[U_k|U_1^1, \ldots, U_{k-1}^1]$. This completes one iteration of the sampler and results in a generated vector $(U_1^1, \ldots, U_k^1)$. Repeating this process, after $i$ iterations we arrive at a generated vector $(U_1^i, \ldots, U_k^i)$. It can be shown that, under mild regularity conditions (see, for example, Geman and Geman, 1984), as $i \to \infty$ this random vector tends in distribution to a random vector having the joint distribution $[U_1, \ldots, U_k]$.

One possible procedure for obtaining summaries of aspects of $[U_1, \ldots, U_k]$ of interest is therefore the following. Run $M$ independent parallel replications of the above sampling procedure, so that, for $i$

judged to be sufficiently large, the resulting generated vectors $(U_{1j}^i,\ldots,U_{kj}^i)$ , $j = 1,2,\ldots,M$ , can be regarded as an *iid* sample of size $M$ from $[U_1,\ldots,U_k]$. Standard moment, quantile or density estimation techniques can then be employed to estimate summary features of interest.

In the Bayesian inference context, $[U_1,\ldots U_k]$ is the joint posterior density of unknown model parameters $U_1,\ldots,U_k$. Univariate marginal inference summaries for $U_s$, say, are simply obtained from $(U_{s1}^i,\ldots,U_{sM}^i)$. Generally, if marginal inference summaries are required for a specified function of $(U_1,\ldots,U_k)$, an *iid* sample of size $M$ from the corresponding marginal density is immediately available on substituting the generated vectors $(U_{1j}^i,\ldots,U_{kj}^i)$, $j = 1,2,\ldots,M$, into the functional form. Exploration of bivariate (or higher dimensional) marginal summaries proceeds in an obvious manner. For further detail, and comments on the pragmatics of choices of $i$ and $M$, see Gelfand and Smith (1990, 1991), Gelfand *et al* (1990, 1991).

The Gibbs sampler thus provides a simulation-based alternative to direct numerical integration methods, and one which depends only on our capacity to generate random variates (reasonably efficiently) from the full conditional densities, $[U_s|U_r,\ r\neq s]$. We shall now look at this latter issue in the context of constrained parameter and censored data problems. Our discussion here will be kept to the minimum necessary to give the reader an appreciation of how the Gibbs sampler achieves crucial simplification. For a much more complete discussion, see Gelfand *et al*, (1991).

## 2.2 MODELS WITH CONSTRAINED PARAMETERS

Suppose a parametric model for data $Y$ involves a $k$-dimensional parameter vector $\theta$, constrained to lie in a subset $S^k$ of $R^k$. For simplicity of exposition, we shall assume here that $S^k$ does not depend on $Y$ (as would be the case, for example, if some components of $\theta$ were truncation parameters: see Gelfand *et al*, 1991). Suppose further that $[Y|\theta]$, $[\theta]$ denote the (unconstrained) forms of likelihood and prior, so that the (constrained) joint posterior for $\theta = (\theta_1,\ldots,\theta_k)$ is given by

$$[\theta|Y] = \frac{[Y|\theta]\ [\theta]}{\int_{S^k} [Y|\theta]\cdot[\theta]}\ ,\quad \theta\in S^k\ .$$

Proceeding by direct numerical integration, we see that there is an immediate problem in calculating the normalizing constant and subsequent problems in performing $(k-1)$-dimensional integrals (over subsets of $S^k$) to obtain marginal density forms.

However, consider now the full conditional forms required for the Gibbs sampler. If $S_i^k(\theta_j,\ j\neq i)$ denotes the cross-section of $S^k$ corresponding to the constraints on $\theta_i$ imposed by $S^k$ for specified values of $\theta_j$, $j\neq i$, we have

$$[\theta_i|Y,\ \theta_j,\ j\neq i] \ \propto\ [Y|\theta]\cdot[\theta] \quad ,\quad \theta_i\in S_i^k(\theta_j,\ j\neq i) \quad .$$

Moreover, the constraints region for $\theta_i$ will typically be an interval, or a union of intervals.

It follows that the typical random variate generation task required for the Gibbs sampler in this case, will simply be that of generating from specified univariate density shapes truncated to intervals. This is a relatively straightforward task: in any case, strikingly easier than high-dimensional numerical integration over complicated constraint volumes.

## 2.3 CENSORED DATA PROBLEMS

Suppose a parametric model for data $Y = (Y_1,\ldots,Y_n)$ involves a $k$-dimensional parameter vector $\theta$, with likelihood defined by

$$[Y|\theta] \ =\ \prod_i [Y_i|\theta] \quad .$$

However, suppose that for $j \geq m > 1$ there exist $V_j$, $W_j$ such that, instead of observing $Y_j$ exactly, we simply observe that $Y_j \in (V_j,\ W_j)$, so that the likelihood is actually given by

$$\prod_{i=1}^{m-1} [Y_i|\theta] \cdot \prod_{j=m}^{n} \int_{V_j}^{W_j} [Y_j|\theta] \quad .$$

(We are here assuming a simple, fully specified censoring process for convenience of exposition. For a more general discussion, see Gelfand *et al*, 1991).

In this case, a moment's reflection reveals that the full conditional forms implied by the above likelihood combined with a prior $[\theta]$ are not, in general, easy forms to sample from. In particular, the

integral terms may have no closed-form analytic expressions, so that standard envelope rejection or ratio-of-uniforms sampling techniques are not readily applicable.

However, suppose we consider $Y' = (Y_m, \ldots, Y_n)$ as additional unknowns, so that the unknown model parameters are $(\theta, Y')$, with the data given by $(Y^*, V, W)$, where $Y^* = (Y_1, \ldots, Y_{m-1})$, $V = (V_m, \ldots, V_n)$ and $W = (W_m, \ldots, W_n)$. Consider now the full conditionals required for the Gibbs sampler:

$$[\theta_i | Y^*, V, W, \theta_j, j \neq i, Y'] \quad , \quad i = 1, \ldots, k \quad ,$$

$$[Y_r | Y^*, V, W, \theta, Y_s, s \neq r, s \geq m] \quad , \quad r = m, \ldots, n \quad .$$

Careful examination of the conditioning variables reveals that the full conditionals for $\theta_1, \ldots, \theta_k$ reduce to

$$[\theta_i | Y, \theta_j, j \neq i] \quad , \quad i = 1, \ldots, k \quad ,$$

the forms that would have obtained in the uncensored case! Typically, these forms present no difficulty for random variate generation.

For the full conditionals for $Y_m, \ldots, Y_n$, examination of the forms reveals that these reduce to

$$[Y_r | \theta] \; / \; \int_{V_r}^{W_r} [Y_r | \theta] \quad , \quad r = m, \ldots, n \quad ;$$

namely, the sampling distributions for the $Y_r$ restricted to the ranges $(V_j, W_j)$. Again, these typically present no difficulty for random variate generation.

The trick of treating censored observations as unknowns in combination with the Gibbs sampler leads to simple Bayesian calculation strategies in otherwise intractable problems (see, also, Tanner and Wong, 1987, for a related manifestation of the idea). In the next section, we illustrate this concretely by detailing the forms of the Gibbs sampler arising in a range of parametric models used in various kinds of survival studies.

## 3. Illustrations For Parametric Survival Models

### 3.1 ORDERED BINOMIAL PARAMETERS

Consider conditionally independent observations $Y_i \sim$ Binomial $(n_i, \theta_i)$, $i = 1,2,\ldots,k$, where it is known that $\theta_1 \leq \theta_2 \leq \ldots \leq \theta_k$ and we seek to make inferences about the $\theta_i$ (or functions, thereof, such as $\theta_{i+1}-\theta_i$ or $(\theta_{i+1}-\theta_i) \,/\, \theta_i$). Problems of this kind arise, for example, in reliability development testing (Smith, 1977; Fard and Dietrich, 1987), where stages $1,\ldots,k$ correspond to successive improvements in reliability.

If the joint prior density is taken to be proportional to

$$\prod_{i=1}^{k} \theta_i^{\alpha_i - 1} (1-\theta_i)^{\beta_i - 1} \quad ,$$

over the simplex, $S^k = \{(\theta_1,\ldots,\theta_k) \;:\; 0 \leq \theta_1 \leq \theta_2 \leq \ldots \leq \theta_k \leq 1\}$, by conjugacy the joint posterior $[\theta|Y]$ has the same form, with support $S^k$, but with $\alpha_i$, $\beta_i$ replaced by $\alpha_i + Y_i$, $\beta_i + n_i - Y_i$, respectively.

Implementation of the Gibbs sampler is now seen to be very simple. The full conditionals are given by

$$[\theta_i|Y,\ \theta_j,\ j \neq i] \;=\; \text{Beta}\,(\alpha_i + Y_i,\ \beta_i + n_i - Y_i) \quad , \quad i = 1,\ldots,k \quad ,$$

restricted to the interval $\theta_{i-1} \leq \theta_i \leq \theta_{i+1}$ ($\theta_0 = 0$, $\theta_{k+1} = 1$), and random variate generation is straightforward.

### 3.2 CENSORED REGRESSION DATA

Schmee and Hahn (1979) modelled log-failure times of motorettes tested at four different temperatures by a straight-line regression of log-failure versus transformed temperature. Censoring occurred whenever a motorette had not failed at the end of the test period. The uncensored case likelihood $[Y|\theta]$ is assumed to derive from $Y_{ij} = \alpha + \beta X_i + \epsilon_{ij}$, where $\epsilon_{ij} \sim N(0,\ \sigma^2)$, $i = 1,\ldots,k$, $j = 1,\ldots,n_i$, but the actual data, $Z$, are given by

$$Z_{ij} \;=\; \begin{cases} Y_{ij} & Y_{ij} \leq W_i \\ \quad \text{if} & \\ W_i & Y_{ij} > W_i \end{cases} \quad ,$$

where $W_i$ is the total test time at temperature corresponding to $X_i$.

To implement the Gibbs sampler, as indicated in Section 2.3, we include the unobserved $Y_{ij}$ (*i.e.*, those where $Y_{ij} > W_i$) as further unknowns in the model, in addition to the basic parameters of interest, $\alpha, \beta$ and $\sigma^2$. Given conjugate normal prior forms for $\alpha, \beta$ and an inverse-gamma prior for $\sigma^2$, it is easily verified that the full conditional forms for $\alpha, \beta$ and $\sigma^2$ are straightforwardly identified conjugate forms (normal, normal and inverse-gamma, respectively) obtained <u>as if</u> all the $Y_{ij}$ were precisely observed. The full conditionals for the unobserved $Y_{ij}$ are simply $N(\alpha + \beta X_i, \sigma^2)$, restricted to the range $Y_{ij} > W_i$. Again, random variate generation from all these full conditionals is unproblematic.

## 3.3 TRUNCATED BIVARIATE NORMAL DATA

Consider a bivariate normal process $(X_i, Y_i)$, $i = 1, \ldots, n$, where some of the $Y_i$ are not observed. One context in which such data arises is in paired survival time studies (using observed logarithms of survival times), where observation $(Y_i)$ of the second of the paired patients is terminated when the first of the pair dies, so that $Y_i$ is observed only if $Y_i \leq X_i$.

More precisely, we assume *iid* pairs $(X_i, Y_i)$ such that for $i = 1, \ldots, n$,

$$\begin{bmatrix} X_i \\ Y_i \end{bmatrix} \sim N \left\{ \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \right\} .$$

We observe the pairs $(X_i, Z_i)$ with $Z_i = Y_i$ if $Y_i \leq X_i$; otherwise we observe $(X_i, *)$, where $Z_i = *$ indicates that $Y_i > X_i$. Suppose that the prior for $(\theta_1, \theta_2)$ is taken to be bivariate normal with mean $(\mu_1, \mu_2)$ and covariance matrix $V$, and that the prior for the covariance matrix, $\Sigma$, say is taken to be an inverse-Wishart, so that $[\Sigma^{-1}] = W\{(\rho R)^{-1}, \rho\}$, with all the hyperparameters $\mu_1$, $\mu_2$, $V$, $\rho$ and $R$ known.

Interest focuses on marginal inferences for $\theta_1$, $\theta_2$ and $\Sigma$, but, following Section 2.3, unobserved values of $Y_i$ are also treated as unknowns in specifying the Gibbs sampler. Defining $T_i = (X_i, Y_i)$, $T =$

$(T_1,\ldots,T_n)$, $\bar{T} = n^{-1}(T_1 + \ldots + T_n)$, $\theta = (\theta_1, \theta_2)$ and $\mu = (\mu_1, \mu_2)$, it is easily verified that

$$[\theta|T, Z, \Sigma] = N\{V(n^{-1}\Sigma + V)^{-1}\,\bar{T} + n^{-1}\Sigma(n^{-1}\Sigma + V)\mu,\ (n\Sigma^{-1} + V^{-1})^{-1}\} \quad ,$$

$$[\Sigma^{-1}|T, Z, \theta] = W\{(\Sigma_i(T_i - \theta)(T_i - \theta)' + \rho R)^{-1},\ n + \rho\}$$

and

$$[Y_i|X_i, Z_i, \theta, \Sigma] = N\{\theta_2 + \sigma_{12}\sigma_1^{-2}(X_i - \theta_1),\ \sigma_2^2 - \sigma_{12}\sigma_1^{-2}\} \quad ,$$

truncated to $(X_i, \infty)$ if $Z_i = *$, with $Y_i$ degenerate at $Z_i$ otherwise. The required random variate generation is routine.

### 3.4 WEIBULL PROPORTIONAL HAZARDS WITH CENSORING

Consider a survival time model in which the hazard function $\lambda(t; Z)$, for an individual with covariate values $Z$ at time $t$, is given by

$$\lambda(t; X) = \rho t^{\rho-1} \exp(Z\beta) \quad ,$$

where $\beta = (\beta_0, \beta_1,\ldots,\beta_p)'$ is a vector of unknown regression parameters and $\rho > 0$ is the unknown Weibull shape parameter.

If $t_1,\ldots,t_n$ are explicitly observed survival times and $t_{n+1},\ldots,t_m$ are censored $(T > t)$ lifetimes, with $Z_j$ denoting covariate values for the $j$*th* case, the likelihood is given by

$$\left[\prod_{j=1}^{n} \rho t_j^{\rho-1}\, e^{Z_j\beta}\right]\left[\prod_{j=1}^{n+m} \exp\left(-t_j^{\rho}\, e^{Z_j\beta}\right)\right] \quad .$$

Clearly, whatever the prior specification, the resulting $(p + 2)$-dimensional posterior is awkward to handle using standard numerical integration procedures.

However, it is easily verified that the second partial derivatives of the log-likelihood with respect to each of the $p + 2$ unknown parameters are all non-positive (see Dellaportas and Smith, 1991). If the prior density is chosen to be log-concave, it follows that all the posterior full conditionals are log-concave. The import of this observation is that highly efficient methods exist for random variate generation from log-concave densities (see, in particular, Gilks and Wild, 1991), so

that routine, straightforward Bayesian calculation for widely used cases of proportional hazards models is possible (see Dellaportas and Smith, 1991, for wider exploitation of log-concavity).

## 4. A Nonparametric Illustration

### 4.1 INTRODUCTION

Nonparametric Bayesian inference for the survival function with right censored data has been studied by Susarla and Van Ryzin (1976), and Ferguson and Phadia (1979). However, we often encounter the situation where some observations are censored from the left and some observations are censored from the right (see Turnbull, 1974, for references to papers addressing doubly censored data sets from a frequentist perspective).

In this section, we study a nonparametric Bayesian approach to such problems, which allows us to incorporate prior beliefs and frees us from making a restrictive (parametric) model assumption for the survival function. Specifically, we assume that the distribution function $F$ of survival times has a prior given by Ferguson's (1973) Dirichlet process, $D(\alpha)$. The measure $\alpha$ can be written as $NF_0$, where $F_0$ is the prior mean of $F$ and $F_0(1 - F_0) / (N + 1)$ is the prior variance of $F$. The larger $N$, the more strongly the prior specifies that $F$ concentrates around $F_0$.

In the doubly censored data case, it is very difficult to obtain an explicit expression for non-parametric Bayesian estimators even in the form of the posterior mean. We shall show, however, that the Gibbs sampler approach, which augments the data by using latent variables that decompose the number of the censored observation into the possible numbers of observations falling into each interval, provides a straightforwardly computed numerical solution. As illustrated in Section 2.3, this augmentation facilitates the specification of appropriate full conditional densities, particularly here for the survival functions given the latent variables. The iterated sampling scheme then allows us to approximate the posterior distribution of the survival function.

### 4.2 THE MODEL

We shall illustrate the approach using a model similar to that studied by Turnbull (1974), who proposed a self-consistent algorithm for computing the generalized maximum likelihood estimators. Here, we add the Dirichlet process prior to the model.

Let $T_1, T_2, \ldots, T_n$ denote the true survival times of $n$ individuals that

could be observed precisely if no censoring were present. The $T_i$ are independent and identically distributed with distribution $F$; that is, $F(t) = P(T \leq t)$ for $t \geq 0$. We consider the case that not all $T_i$ are observed precisely. For each $i$, we assume that there are "windows" of observations $V_i$ and $W_i$ ($V_i \leq W_i$) that are either fixed constants or random variables independent of the $\{T_i\}$. We observe

$$X_i = \max [\ \min(T_i, W_i), V_i] \quad .$$

Moreover, for each item, we also know whether it is left-censored with $X_i = V_i$, or right-censored with $X_i = W_i$, or a precisely observed time with $X_i = T_i$.

We assume that items (or patients) are examined at discrete times (for example, monthly) and that there is a natural discrete time scale $0 < t_1 < t_2 < \ldots, t_m$, with observed deaths classified into one of the $m$ intervals $(0, t_1], (t_1, t_2] \ldots, (t_{m-1}, t_m]$. Let $\delta_i$ denote the number of precise observations (=) in the period $(t_{i-1}, t_i]$, $\mu_i$ denote the number of left-censored (≤) entries at age $t_i$, and $\lambda_i$ denote the number of right-censored (>) entries at $t_i$. It is assumed that the left-censored entries $\mu_i$ all occur at the end of age period $(t_i, t_{i+1}]$. The data can then be summarized by the following tabulation:

| Type of obs. \ age | $(0, t_1]$ | $(t_1, t_2]$ | ... | $(t_{m-1}, t_m]$ |
|---|---|---|---|---|
| (=) | $\delta_1$ | $\delta_2$ | ... | $\delta_m$ |
| (≤) | $\mu_1$ | $\mu_2$ | ... | $\mu_m$ |
| (>) | $\lambda_1$ | $\lambda_2$ | ... | $\lambda_m$ |

Let $P_j = P(t_j) = 1 - F(t_j)$ denote the survival function evaluated at $t_j$, so that the likelihood function is proportional to

$$\prod_{j=1}^{m} (P_{j-1} - P_j)^{\delta_j} (1 - P_j)^{\mu_j} P_j^{\lambda_j} \quad .$$

Let $\theta_j = P_{j-1} - P_j$ for $j = 1, \ldots, m$ and let $\theta_{m+1} = P_m$. The prior process specifies that the distribution of the $\theta$'s is the Dirichlet distribution

$$\pi(\theta) = C \prod_{j=1}^{m+1} (\theta_j)^{\alpha_j - 1} ,$$

where

$$\alpha_j = N(F_0(t_j) - F_0(t_{j-1})) ,$$

for $j = 1, \ldots, m+1$, with $F_0(t_{m+1}) = 1$, and

$$C = \frac{\Gamma(N)}{\prod_{j=1}^{m+1} \Gamma(\alpha_j)} .$$

The posterior distribution of $\theta = (\theta_1, \theta_2, \ldots, \theta_m, \theta_{m+1})$ is known to be a mixture of Dirichlet distributions (see Antoniak, 1974). In the next section, we show how the Gibbs sampler side-steps the need for direct computation of this mixture.

## 4.3 APPROXIMATION VIA THE GIBBS SAMPLER

To employ the Gibbs sampler, we use the idea of Section 2.3 and introduce latent variables that decompose the numbers of censored entries into the numbers of observations belonging to individual intervals. Let $Z_{1j}, Z_{2j}, \ldots, Z_{jj}$ denote the random variables that count the number of observations in $\mu_j$ that might fall in the intervals $(0, t_1], (t_1, t_2], \ldots, (t_{j-1}, t_j]$, respectively, so that $\mu_j = \sum_{l=1}^{j} Z_{lj}$. Further, let $Z_{j+1\,j}, \ldots, Z_{m+1\,j}$ denote the number of observations in $\lambda_j$ that might fall in the intervals $(t_j, t_{j+1}], \ldots, (t_{m-1}, t_m], (t_m, \infty]$, respectively, so that $\lambda_j = \sum_{l=j+1}^{m+1} Z_{lj}$.

Our objective is to summarize, via samples generated form the Gibbs sampler, the posterior distribution of $\theta$ given the data. The posterior full conditional for $\theta$ given the Z's and the data, is easily seen to be an up-dated Dirichlet distribution depending only on the Z's. The posterior full conditional for the Z's given $\theta$ and the data, is easily seen to be a product of multinomial distributions. Thus, suppose at the $i$th iteration step of the Gibbs sampler, we have the realization $\theta^i = (\theta_1^i, \theta_2^i, \ldots, \theta_{m+1}^i)$, with $\sum_{l=1}^{m+1} \theta_l^i = 1$. We then up-date the Z variables from the multinomial distributions as follows. For each $j$, $j = 1, \ldots, m$, we sample $Z_{1j}^{i+1}, \ldots, Z_{jj}^{i+1}$ from the multinomial distribu-

tion with sample size $\mu_j$ and parameters $r^i_{1j},\ldots,r^i_{jj}$, where $r^i_{lj} = \theta^i_l / \sum_{l=1}^{j} \theta^i_l$ for $l = 1,\ldots,j$. Similarly, we sample $Z^{i+1}_{j+1j},\ldots,Z^{i+1}_{m+1j}$ from the multinomial distribution with sample size $\lambda_j$ and parameters $r^i_{j+1j},\ldots,r^i_{m+1j}$, where $r^i_{lj} = \theta^i_l / \sum_{l=j+1}^{m+1} \theta^i_l$ for $l = j+1,\ldots,m+1$. Having sampled the $Z$ random variables, we then generate new $\theta$ variables from the Dirichlet distribution as follows. We compute, for each $l$, $l = 1,\ldots,m+1$,

$$Y^{i+1}_l = \alpha_l + \delta_l + \sum_{j=1}^{m} Z^{i+1}_{lj} \quad ,$$

and then sample $(\theta^{i+1}_1,\ldots,\theta^{i+1}_m, \theta^{i+1}_{m+1})$ from the Dirichlet distribution with parameters $(Y^{i+1}_1,\ldots,Y^{i+1}_{m+1})$.

By running $M$ parallel independent replications of the sampler, after the $i$th iteration, we have $\theta^i_{1s}, \theta^i_{2s},\ldots,\theta^i_{m+1,s}$, and $Y^i_{1s},\ldots,Y^i_{m+1,s}$, for $s = 1,..,M$. The posterior distribution of $\theta_l$ for $l = 1,\ldots,m+1$ can then be approximated (for sufficiently large $i$) by

$$\hat{F}(\theta_l|data) = M^{-1} \sum_{s=1}^{M} Beta(Y^i_{ls}, \sum_{k\neq l}^{m+1} Y^i_{ks}) \quad ,$$

where $Beta(a,b)$ denotes the beta density with parameters $a$ and $b$. A posterior estimate of the $\theta_l$ is then given by

$$\hat{\theta}_l = M^{-1} \sum_{s=1}^{M} \frac{Y^i_{ls}}{\sum_{l=1}^{m+1} Y^i_{ls}} \quad .$$

Other posterior summaries can be computed similarly from the replicated samples, $i$ and $M$ having been selected to achieve "convergence" to "smooth" estimates.

## 4.4 A NUMERICAL EXAMPLE

To illustrate the Gibbs sampler technique, we shall reanalyze the data set given by Kaplan and Meier (1958). The data consist of deaths occurring at .8, 3.1, 5.4 and 9.2 months and losses occurring at 1.0, 2.7, 7.0 and 12.1 months. For comparison purposes, we consider the same prior specifications used by Susarla and Van Ryzin (1976) in their

Bayesian reanalysis of the data. That is, $F_0(t) = 1 - e^{-\phi t}$ with $\phi$ = .12 and $N$ = 4, 8, and 16.

To apply the Gibbs sampler approach, we divide the positive real line into the following intervals: $(0, .8^-]$, $(.8^-, .8]$, $(.8, 1]$, $(1, 2.7]$, $(2.7, 3.1^-]$, $(3.1^-, 3.1]$, $(3.1, 5.4^-]$, $(5.4^-, 5.4]$, $(5.4, 7]$, $(7, 9.2^-]$, $(9.2^-, 9.2]$, $(9.2, 12.1]$, and $(12.1, \infty)$. We label these intervals by $(0, t_1]$, $(t_1, t_2], \ldots, (t_{12}, t_{13}]$, and let $\theta_1$, $\theta_2, \ldots,$ and $\theta_{13}$, respectively, denote the probabilities assigned to the intervals.

The likelihood of $\theta$ is proportional to

$$L(\theta) = \theta_2\theta_6\theta_8\theta_{11}(\theta_4 + \theta_5 + \ldots + \theta_{13}) \times (\theta_5 + \ldots + \theta_{13})(\theta_{10} + \ldots + \theta_{13})\theta_{13} .$$

Let $\alpha_l = N(e^{-\phi t_{l-1}} - e^{-\phi t_l})$, so that the prior distribution of $\theta$ is

$$\pi(\theta) = C \prod_{l=1}^{13} \theta_l^{\alpha_l - 1} ,$$

where $C$ is the normalizing constant.

Note that $\theta_2$, $\theta_6$, $\theta_8$, $\theta_{11}$ and $\theta_{13}$ in the likelihood combine simply with the corresponding $\theta$ variables in the prior distribution, so that the parameters $\theta_2$, $\theta_6$, $\theta_8$, $\theta_{11}$ and $\theta_{13}$ are each up-dated by 1 in the posterior distribution. Therefore, we need only introduce three $Z$ variables for the incomplete data, namely, $Z_1 = (Z_{41}, Z_{51}, \ldots, Z_{13,1})$, $Z_2 = (Z_{52}, Z_{62}, \ldots, Z_{13,2})$, and $Z_3 = (Z_{10,3}, Z_{11,3}, Z_{12,3}, Z_{13,3})$. We then sample $Z_j$, for $j$ = 1, 2, and 3, from the appropriate multinomial distribution with sample size 1 and rescaled probabilities.

To estimate the survival function at $t_j$, we accumulate the $\theta_l$ for $l > j$. For $t$ between $t_j$ and $t_{j+1}$, an interpolation formula that connects the survival function at the two end points according to the prior shape can be used. Tables 1 and 2 exhibit the Gibbs sampler results for the survival function evaluated at $t_j$ with $M$ = 1000 and $M$ = 4000, both with $I$ = 10. The exact Bayes solutions given by Susarla and Van Ryzin are also listed for comparison. The tables show that the Gibbs sampler results for $M$ = 1000 are already very accurate in approxima-

ting the exact Bayes rules. Similar results hold for $N = 16$. For further illustration of the Gibbs sampler methodology, see Kuo (1991), who reanalyses data from Turnbull (1974).

Table 1: Gibbs Approximation to the Bayes Estimates for $N = 4$

| Statistics \ age($t$) | $.8^-$ | .8 | 1 | 2.7 | $3.1^-$ | 3.1 |
|---|---|---|---|---|---|---|
| $\hat{P}_t$ with $M = 1000$ | .970 | .886 | .879 | .819 | .805 | .702 |
| $\hat{P}_t$ with $M = 4000$ | .970 | .886 | .879 | .819 | .805 | .701 |
| Exact Bayes | .970 | .886 | .879 | .819 | .805 | .701 |

| Statistics \ age($t$) | $5.4^-$ | 5.4 | 7 | $9.2^-$ | 9.2 | 12.1 |
|---|---|---|---|---|---|---|
| $\hat{P}_t$ with $M = 1000$ | .632 | .529 | .491 | .437 | .305 | .253 |
| $\hat{P}_t$ with $M = 4000$ | .632 | .529 | .491 | .438 | .307 | .256 |
| Exact Bayes | .632 | .528 | .490 | .438 | .306 | .255 |

Table 2: Gibbs Approximation to the Bayes Estimates for $N = 8$

| Statistics \ age($t$) | .8- | .8 | 1 | 2.7 | $3.1^-$ | 3.1 |
|---|---|---|---|---|---|---|
| $\hat{P}_t$ with $M = 1000$ | .954 | .892 | .881 | .792 | .773 | .698 |
| $\hat{P}_t$ with $M = 4000$ | .954 | .892 | .881 | .792 | .773 | .700 |
| Exact Bayes | .954 | .892 | .881 | .793 | .773 | .699 |

| Statistics \ age($t$) | $5.4^-$ | 5.4 | 7 | $9.2^-$ | 9.2 | 12.1 |
|---|---|---|---|---|---|---|
| $\hat{P}_t$ with $M = 1000$ | .600 | .527 | .474 | .405 | .316 | .249 |
| $\hat{P}_t$ with $M = 4000$ | .602 | .529 | .474 | .405 | .318 | .250 |
| Exact Bayes | .602 | .528 | .474 | .405 | .318 | .250 |

ACKNOWLEDGEMENTS

The work of the first author is supported by NSF grant DMS9008021 and the Naval Postgraduate School, Monterey, CA 93943, U. S. A. Both authors have benefited from discussions with A. E. Gelfand.

REFERENCES

ANTONIAK, C. E., (1974). "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems". *Ann. Statist.*, 2, 1152-74.

DELLAPORTAS, P. and SMITH, A. F. M., (1991). "Bayesian inference for Generalised Linear and Proportional Hazards Models via Gibbs Sampling". *Applied Statistics*. To appear.

FARD, N. S. and DIETRICH, D. L., (1987). "A Bayesian note on reliability growth during a development testing programme". *I. E. E. E. Trans. Rel.*, R-36, 568-572.

FERGUSON, T. S., (1973). "A Bayesian analysis of some nonparametric problems". *Ann. Statist.*, 1, 209-30.

FERGUSON, T. S. and PHADIA, E. G., (1979). "Bayesian nonparametric estimation based on censored data". *Ann. Statist.*, 7, 163-86.

GELFAND, A. E. and SMITH, A. F. M., (1990). "Sampling-based approaches to calculating marginal densities". *J. Am. Statist. Assoc.*, 85, 398-409.

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. Amer. Statist. Assoc.*, 85, 972-985.

GELFAND, A. E. and SMITH, A. F. M. (1991). "Gibbs sampling for marginal posterior expectations". *Commun. in Statist.* To appear.

GELFAND, A. E., SMITH, A. F. M. and LEE, T-M., (1991). "Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling". *J. Am. Statist. Assoc.* To appear.

GEMAN, S. and GEMAN, D., (1984). "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images". *I. E. E. E. Trans. on Pattern Analysis and Machine Intelligence*, 6, 721-741.

GILKS, W. R. and WILD, P.,(1991). "Adaptive rejection sampling for Gibbs Sampling". *Applied Statistics*. To appear.

KAPLAN, E. L. and MEIER, P., (1958). "Nonparametric estimation from incomplete observations." *J. Am. Statist. Assoc.*, 53, 457-81.

KUO, L. (1991). "Sampling based approach to computing nonparametric Bayesian estimators with doubly censored data". *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface.*

SCHMEE, J. and HAHN, G. J., (1979). "A simple method for regression analysis with censored data". *Technometrics*, **21**, 417-432.

SMITH, A. F. M., (1977). "A Bayesian note on reliability growth during a development testing programme". *I. E. E. E. Trans. Rel.*, **R-26**, 346-347.

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". *The Statistician*, **36**, 75-82.

SUSARLA, V. and VAN RYZIN, J., (1976). "Nonparametric estimation of survival curves from incomplete observations." *J. Am. Statist. Assoc.*, **71**, 897-902.

TANNER, M. and WONG, W., (1987). "The calculation of posterior distributions by data augmentation", (with discussion). *J. Am. Statist. Assoc.*, **81**, 82-86.

TURNBULL, B. W., (1974). "Nonparametric estimation of a survivorship function with doubly censored data." *J. Am Statist. Assoc.*, **69**, 169-73.

[FILE: ASBCSM.CHI (JUNE 1991)]

## INITIAL DISTRIBUTION LIST

1. Library (Code 0142) ..... 2
   Naval Postgraduate School
   Monterey, CA 93943-5000

2. Defense Technical Information Center ..... 2
   Cameron Station
   Alexandria, VA 22314

3. Office of Research Administration (Code 012) ..... 1
   Naval Postgraduate School
   Monterey, CA 93943-5000

4. Prof. Peter Purdue ..... 1
   Code OR-Pd
   Naval Postgraduate School
   Monterey, CA 93943-5000

5. Department of Operations Research (Code 55) ..... 1
   Naval Postgraduate School
   Monterey, CA 93943-5000

6. Prof. James D. Esary ..... 1
   Code OR/Ey
   Naval Postgraduate School
   Monterey, CA 93943-5000

7. Prof. Donald Gaver, Code OR-Gv ..... 1
   Naval Postgraduate School
   Monterey, CA 93943-5000

8. Prof. Patricia Jacobs ..... 1
   Code OR/Jc
   Naval Postgraduate School
   Monterey, CA 93943-5000

9. Prof. Harold J. Larson ..... 1
   Code OR/La
   Naval Postgraduate School
   Monterey, CA 93943-5000

10. Prof. P. A. W. Lewis ..... 1
Code OR/Lw
Naval Postgraduate School
Monterey, CA 93943-5000

11. Prof. Paul R. Milch ..... 1
Code OR/Mh
Naval Postgraduate School
Monterey, CA 93943-5000

12. Prof. Robert R. Read ..... 1
Code OR/Re
Naval Postgraduate School
Monterey, CA 93943-5000

13. Prof. Lyn R. Whitaker ..... 1
Code OR/Wh
Naval Postgraduate School
Monterey, CA 93943-5000

14. Prof. W. Max Woods ..... 1
Code OR/Wo
Naval Postgraduate School
Monterey, CA 93943-5000

15. Prof. Thomas Ferguson ..... 1
Math Department
University of California
Los Angeles, CA 90024

16. Prof. Robert Jennrich ..... 1
Math Department
University of California
Los Angeles, CA 90024

17. Prof. Keh-Chau Li ..... 1
Math Department
University of California
Los Angeles, CA 90024

18. Prof. Donald Ylvisker ..... 1
Math Department
University of California
Los Angeles, CA 90024

19. Prof. Barry Arnold ..... 1
Statistics Department
University of California
Riverside, CA 92521

20. Prof. Keh-Shin Lii ..... 1
Statistics Department
University of California
Riverside, CA 92521

21. Prof. James Press ..... 1
Statistics Department
University of California
Riverside, CA 92521

22. Prof. David Strauss ..... 1
Statistics Department
University of California
Riverside, CA 92521

23. Prof. Joseph R. Gani ..... 1
Mathematics Department
University of California
Santa Barbara, CA 93106

24. Prof. S. R. Jammalamadaka ..... 1
Statistics Department
University of California
Santa Barbara, CA 931067

25. Prof. John Hsu ..... 1
Statistics Department
University of California
Santa Barbara, CA 931067

26. Prof. Milton Sobel ..... 1
Statistics Department
University of California
Santa Barbara, CA 931067

27. Prof. Bradley Efron ..... 1
Statistics Dept.
Sequoia Hall
Stanford University
Stanford, CA 94305

28. Prof. Iain Johnstone ..... 1
Statistics Department
Stanford University
Stanford, CA 94305

29. Prof. Herbert Solomon ..... 1
Statistics Department
Stanford University
Stanford, CA 94305

30. Prof. Byron Brown ..... 1
Biostatistics Department
Stanford University
Stanford, CA 94305

31. Prof. George Roussas ..... 1
Statistics Division
University of California
Davis, CA 95616

32. Prof. P. K. Bhattacharya ..... 1
Statistics Division
University of California
Davis, CA 95616

33. Prof. Frank Samaniego ..... 1
Statistics Division
University of California
Davis, CA 95616

34. Prof. Jane-Ling Wang ..... 1
Statistics Division
University of California
Davis, CA 95616

35. Dr. Jack C. Lee ..... 1
Bell Communications Research
Morris Research and Engineering Center
435 South Street, 2Q-374
Morristown, NJ 07960

36. Prof. Malay Ghosh ..... 1
Statistics Department
University of Florida
Cainvesville, FL 32611

37. Prof. Glen Meeden ..... 1
Theoretical Statistics Department
University of Minnesota
270 Vincent
206 Church St. SE
Minneapolis, MN 55455

38. Prof. James Zidek ..... 1
Statistics Department
University of British Columbia
Vancouver, V6T 1W5 CANADA

39. Prof. Bruce Turnbull ..... 1
Operations Research and Industrial Engineering
Cornell University
227 FfTC
Ithaca, NY 14853-3801

40. Center for Naval Analyses ..... 1
4401 Ford Avenue
Alexandria, VA 22302-0268

41. Dr. Terry Speed ..... 1
Statistics Department
University of California
Berkeley, CA 94720

42. Dr. P. Welch ..... 1
IBM Research Laboratory
Yorktown Heights, NY 10598

43. Dr. Neil Gerr ..... 1
Office of Naval Research
Arlington, VA 22217

44. Dr. J. Abrahams, Code 1111, Room 607 ..... 1
Mathematical Sciences Division, Office of Naval Research
800 North Quincy Street
Arlington, VA 22217-5000

45. Prof. H. Chernoff ..... 1
Department of Statistics
Harvard University
1 Oxford Street
Cambridge, MA 02138

46. Prof. Carl N. Morris ..... 1
Statistics Department
Harvard University
1 Oxford St.
Cambridge, MA 02138

47. Prof. Tom A. Louis ..... 1
School of Public Health
University of Minnesota
Mayo Bldg. A460
Minneapolis, MN 55455

48. Prof. William Jewell ..... 1
Operations Research Department
University of California, Berkeley
Berkeley, CA 94720

49. Dr. S. R. Dalal ..... 1
Bellcore
445 South Street
Morristown, NJ 07962-1910

50. Dr. Michael P. Cohen ..... 1
National Center for Education Statistics
555 New Jersey Ave., NW
Washington, DC 20208-5654

51. Prof. Lynn Kuo ..... 1
Code OR-Ku
Naval Postgraduate School
Monterey, CA 93943-5000