# BAYESIAN NONPARAMETRIC SURVIVAL ANALYSIS

by

Lin Yuan

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo. Ontario. Canada. 1997

The University of Waterloo requires the signatures of all persons using or photocopying this thesis. Please sign below, and give address and date.

# Acknowledgements

First, I wish to express my sincere gratitude to my supervisor John D. Kalbfleisch for his expertise, guidance and enthusiasm in the research of Bayesian nonparametric methodology.

Thanks are due to my Committee: Mary Thompson, Don McLeish, David Matthews, Kjell Doksum and Andrew Heunis for their helpful suggestions and comments. I am especially grateful to Professor Matthews for his contribution that improved the presentation of this thesis.

Last, but most importantly, I thank my family and friends that have supported and encouraged me during the years of diligent study at Waterloo.

*To My Parents*

# *Abstract*

This thesis makes contributions to the Bayesian nonparametric approach for survival and bioassay problems. It contains creative work towards a simple and practical Bayesian analysis for right-censored failure time data using a smoothed prior, and for binary and doubly-censored data using the Dirichlet process prior.

One-sample survival analysis under a smoothed prior is fully studied. The posterior computations are realized via the Gibbs sampler, and illustrated by numerical examples. Bayesian inference under non-informative priors is addressed and compared with existing results. A compromised version of Bayesian nonparametric approach is proposed which retreats from the infinite-dimensional priors and considers a more practical treatment using data-dependent priors. Links to some well-known results such as Cox's partial likelihood for proportional hazards regression and Hill's rule for prediction are established. Fiducial inference for failure time data is also discussed, which is numerically equivalent to the Bayesian approach under a non-informative and data-dependent prior.

A new auxiliary variables technique is proposed which has substantially simplified the Bayesian bioassay under a Dirichlet process prior, and application is illustrated in cancer risk assessment. The problem of combining many assays is discussed in the empirical Bayes framework, and more complicated types of data such as doubly-censored data are also considered.

# *Contents*

*CHAPTER 1*

# *Introduction and Review*

## 1.1 Introduction

Bayesian nonparametric statistics has enjoyed limited success since the fundamental work of Ferguson (1973). The current status of this branch of Bayesian statistics is better described as a research topic rather than a well developed theory and application tool. The nonparametric approach makes few model assumptions yet incorporates initial information, but sophisticated posterior characterization and costly computation are often inevitable. Therefore, the basic issue is not philosophical but rather technical.

Ferguson (1973) generalized the traditional Dirichlet distribution to the Dirichlet process, which stimulated a series of investigations in this particular area. As is well known, the Dirichlet process prior facilitates simple calculation and leads to many natural results. On the other hand, the Dirichlet process prior assigns full probability mass to the class of discrete distributions and this has caused undesirable sampling properties, a lack of smoothness in results and inconvenience in applications. Despite some criticisms, Ferguson's prior has received much attention for its computational convenience. Many have followed his work and attempted to repair its defects. For example, Susarla and Van Ryzin (1976) applied the Dirichlet process to survival analysis, and Lo (1984) discussed the smoothing problem of the discrete estimate.

Neutral to the right processes, another class of priors, were introduced by Doksum (1974). These place full probability mass on discrete life distributions and yield tractable posteriors for right-censored data. Applications are mainly in survival analysis; for example, Kalbfleisch

(1978) analyzed the proportional hazards model using the Gamma process, a special neutral to the right process.

How far can we go with the Bayesian nonparametric methods? This is determined by the inherent structure of Bayesian inference. As is well-known, the Bayesian posterior is essentially a product of the prior and the likelihood. Therefore, added complexity in the prior or the likelihood results in a more complex posterior. Although the Dirichlet and neutral to the right process priors lead to tractable Bayesian analysis for complete and right-censored observations, difficulty may arise when more features are present either in the prior or in the data. For example, if we add smoothness to the prior, or we have more complicated data such as doubly-censored data, the posterior computation generally becomes difficult.

There have been efforts to find suitable smoothed priors. For example, a prior was constructed by Dykstra and Laud (1981) on continuous survival functions with an increasing hazard function. For the purpose of density estimation, the so-called logistic-Gaussian prior was proposed by Leonard (1978), and further studied by Lenk (1988, 1991). Unfortunately, all these suffered from the lack of a proper device for a full posterior computation.

In recent years, powerful numerical devices have been developed to deal with high-dimensional posteriors. Resampling methods, especially the Gibbs sampler, have made many difficult computations possible. A question arises as to whether the Gibbs sampler can provide a way to solve the nonparametric problem using a smoothed prior. One must understand that, however, the so-called parameter in the nonparametric set-up is usually of infinite dimension, a situation where Gibbs sampler cannot be applied directly.

Some progress has been made using a kind of hierarchical model introduced by Escobar (1994) and Escobar and West (1995) that puts the Dirichlet process prior on the hyper-parameter. This allows a finite-dimensional posterior analysis for the parameters since the truly nonparametric part is in the background. The drawback is that a Gibbs sampler has to be run in its original form, which is highly iterative, and over $n$ parameters if there are $n$ observations. This is rather costly unless the sample size is small.

Complexity in data also poses problems in posterior computation. Even the Dirichlet process prior leads to analytically intractable results for binary data. For example, a Bayesian bioassay under a mixture of Dirichlet process priors (Antoniak, 1974) is hard to implement; a more realistic treatment (Gelfand and Kuo, 1991) is provided by the Gibbs sampler.

This thesis is aimed at extending the existing theory in both directions: (1) smoothness in the prior and (2) more complicated types of data. Specifically, it facilitates Bayesian analysis for right-censored failure time data using smoothed priors, and for binary and doubly-censored data using Dirichlet process priors.

## 1.2 The Dirichlet Process Prior

The Dirichlet distribution is well known to statisticians. Its basic properties can be found in textbooks such as Wilks (1962). A $k$-parameter Dirichlet distribution $D(\alpha_1, \cdots, \alpha_k)$, where $\alpha_i > 0$, is a probability distribution confined to a $(k-1)$-dimensional manifold with density function

$$f(u_1, \cdots, u_{k-1}) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} u_1^{\alpha_1 - 1} \cdots u_{k-1}^{\alpha_{k-1} - 1} (1 - u_1 - \cdots - u_{k-1})^{\alpha_k - 1}$$

for $u_i > 0$, and $u_1 + \cdots + u_{k-1} < 1$, otherwise it is zero. When $s = 2$, $D(\alpha_1, \alpha_2)$ is just the Beta distribution.

Let $\alpha = \alpha_1 + \cdots + \alpha_k$ and $\pi_i = \alpha_i / \alpha$, then the Dirichlet distribution can be written as $D(\alpha\pi_1, \cdots, \alpha\pi_k)$, where $\alpha > 0$ is called the confidence parameter, and $(\pi_1, \cdots, \pi_k)$, which is also a probability distribution, is called the shape parameter. If the distribution of a random vector $(p_1, \cdots, p_k)$ is the Dirichlet $D(\alpha_1, \cdots, \alpha_k)$, then the marginal distribution of $p_i$ is $Beta(\alpha_i, \alpha - \alpha_i)$ and the joint distribution of $p_i$ and $p_j$ is $D(\alpha_i, \alpha_j, \alpha - \alpha_i - \alpha_j)$.

The Dirichlet distribution is a conjugate prior for multinomial models. Suppose we have a prior guess at the unknown distribution, say $(\pi_1, \cdots, \pi_k)$. Then, the prior distribution $D(c\pi_1, \cdots, c\pi_k)$ is recommended because the prior mean of each $p_i$ is $\pi_i$ which coincides with

our guess. The parameter $c$ is a measure of our confidence about the guess, where a larger value of $c$ implies more concentration of the prior around $(\pi_1, \cdots, \pi_k)$.

As a natural generalization of the Dirichlet distribution, the Dirichlet process introduced by Ferguson (1973) is a random measure on an abstract space. It can be briefly described as follows. Let $\alpha$ be a finite measure with positive total mass on measurable space $(\mathcal{X}, \mathcal{A})$. Then a random measure $P$ or, equivalently, a random probability function $F$ induced by $P$ is said to be a Dirichlet process with parameter $\alpha$ if, for any partition of the whole space $A_1, \cdots, A_k$ the joint distribution of $P(A_1), \cdots, P(A_k)$ is $D(\alpha(A_1), \cdots, \alpha(A_k))$. In this case, $F_0(A) = \alpha(A)/\alpha(\mathcal{X})$, $A \in \mathcal{A}$ is the shape parameter representing the initial estimate of a distribution, and $c = \alpha(\mathcal{X})$ is the confidence.

For the Dirichlet process on the real line, the distribution of the $q$-th quantile $\xi_q$ is expressed as

$$\Pr(\xi_q \leq t) = 1 - B(q|\alpha(-\infty, t]), \alpha((t, \infty)))$$

where $B(x|\alpha, \beta)$ denote the cumulative probability function of the Beta distribution with parameters $\alpha$ and $\beta$.

The main result in Ferguson (1973) states that, if $X_1, \cdots, X_n$ is a sample from $P$, then the posterior distribution of $P$ given $X_1, \cdots, X_n$ is also a Dirichlet process with parameter $\alpha + \sum_{i=1}^{n} \delta_{X_i}$, where $\delta_x$ denotes the measure giving mass one to $x$. Ferguson has considered many applications including the estimation of a distribution, a mean, a variance and quantiles. Under quadratic loss, the Bayesian estimate of the cumulative distribution function is

$$\hat{F} = \frac{c}{c+n} F_0 + \frac{n}{c+n} F_n$$

where $F_0$ is the initial estimate of the unknown cumulative distribution and $F_n$ is the empirical distribution function. This gives a clear picture of the role of the prior and the data in Bayesian inference.

## 1.3 The Gamma Process Prior

Gamma distributions have been widely used in Bayesian analysis. Let $G(a, b)$ denote the Gamma distribution with shape $a$ and scale $b$ and $P(\theta)$ the Poisson distribution with mean $\theta$. We briefly review the Poisson-Gamma model. Suppose the unknown parameter $\theta$ lies in $(0, \infty)$ and we expect $\theta$ to be near $\theta_0$ with a certain degree of confidence. A conjugate prior would be $G(c\theta_0, c)$ where $\theta_0$ is our initial belief or guess and $c$ is a measure of our confidence in that guess. For a fixed experiment, when $c$ is larger the prior would be more concentrated around $\theta_0$ and contribute more to the posterior. On the other hand, when $c$ is smaller the prior would have less influence in the statistical conclusions. Extremely high confidence happens when $c \to \infty$ which means the prior is degenerating and we are certain about our guess. On the other hand, when $c \to 0$ the prior becomes almost uniform on $\log \theta$.

The Gamma process is an independent increments process with Gamma distributed increments. Its sample paths are non-decreasing pure jump functions. Physical applications can be found in Moran (1959) where the process of inputs to a dam over a time period was modeled as a Gamma process. For Bayesians, such a process can also serve as a subjective probability representing knowledge or uncertainty. Kalbfleisch (1978) created the so-called Gamma process prior for a Bayesian analysis of proportional hazards regression. We start with a guess $\Lambda_0$ about the true cumulative hazard, and assume that the increment $\Lambda(t) - \Lambda(s)$ has a distribution $G(c[\Lambda_0(t) - \Lambda_0(s)], c)$ where $c > 0$ is the degree of belief attached to that guess. As in the Dirichlet process, $\Lambda_0$ and $c$ are identified as shape and confidence parameters.

Suppose each subject with covariates $z = (z_1, \cdots, z_k)'$ has hazard function

$$\lambda(t|z) = \exp(\beta' z)\lambda^*(t), \qquad t \geq 0,$$

where $\beta = (\beta_1, \cdots, \beta_k)'$ are the regression parameters and $\lambda^*(t)$ is the baseline hazard. Let the observed failures be $t_1 < \cdots < t_n$ and suppose censorings in $[t_i, t_{i+1})$ are adjusted to $t_i$. The interest in this case is the estimation of the regression parameter. Kalbfleisch

(1978) assumed that prior knowledge about the baseline hazard could be represented by the Gamma process prior described above. Suppose the subject who failed at $t_i$ has covariate $z_i$ and let $s_i(\beta) = \sum_{j \in \mathcal{R}(t_i)} e^{\beta' z_j}$ where $\mathcal{R}(t_i)$ is the risk set just before $t_i$. Let $B_i = -\log[1 - \exp(\beta' z_i)/(c + s_i(\beta))]$. Then a Bayesian way of eliminating nuisance parameters is adopted and a likelihood for $\beta$

$$L(\beta) = c^n \exp[-\sum_{i=1}^{n} cB_i\Lambda_0(t_i)] \prod_{i=1}^{n} [\lambda_0(t_i)B_i]$$

is obtained by integrating out the baseline hazard. This gives a spectrum of likelihoods ranging from the truly nonparametric situation, where the baseline hazard is completely unknown, to the parametric situation, where the baseline hazard is known.

Given $\beta$, the posterior of the cumulative hazard function is again an independent increments process. Between $t_{i-1}$ and $t_i$ the cumulative hazard function is a Gamma process with shape $c\Lambda_0/(c + s_i(\beta))$ and confidence $c + s_i(\beta)$: at $t_i$ the increment has a density function proportional to

$$u^{-1} \exp(-cu)[\exp(-s_i(\beta)u) - \exp(-s_{i+1}(\beta)u)], \quad u > 0.$$

## 1.4 The Gaussian Process Prior

It seems that the Gaussian process is not suitable for the purpose of assigning a prior to distributions. This is mainly due to the constraints that have to be satisfied by a distribution or density function. Leonard (1978) considered the logistic transform of a Gaussian process. Let $x$ be a Gaussian process on a finite interval $[a, b]$. Then,

$$f(t) = \frac{\exp[-x(t)]}{\int_a^b \exp[-x(s)ds]}$$

is obviously a density function. Leonard argued that the mean and covariance of $x$ can bring the prior information into $f$. Nevertheless, the prior features of $f$, such as prior mean and

variance, are extremely hard to calculate. We might therefore have difficulty in specifying $x$ even if we have knowledge about $f$. The posterior distribution of $f$ given a sample from it is, according to Leonard, rather complicated and thus omitted in his paper. Lenk (1988, 1991) studied the same model but was also unable to make any real progress in posterior computation.

However, the Gaussian process does offer, at least theoretically, the possibility of nice results incorporating initial information about an arbitrary curve. A simple mathematical treatment can be described as follows. Let us confine the curve $x(t)$ to $L^2[a,b]$, the space of square integrable functions with a complete orthonormal basis $\{\psi_n\}$. According to functional analysis, $x(t) = \sum_{n=1}^{\infty} \beta_n \psi_n(t)$, where $\beta_n$ is the $n$-th Fourier coefficient of $x$ and the equality is in the sense of the $L^2$ norm. Initial knowledge about $x$ can be incorporated by assigning a joint distribution on $\{\beta_n\}_{n=1}^{\infty}$. It is simple to assume that $\beta_n \sim N(\mu_n, \lambda_n)$ is an independent sequence satisfying $\lambda_n \geq 0$, $\sum_{n=1}^{\infty} \mu_n^2 < \infty$ and $\sum_{n=1}^{\infty} \lambda_n < \infty$. By doing so, $x(t)$ becomes a Gaussian process on $[a,b]$ with mean

$$\mu_0(t) = \sum_{n=1}^{\infty} \mu_n \psi_n(t)$$

and covariance

$$\gamma_0(s,t) = \sum_{n=1}^{\infty} \lambda_n \psi_n(s) \psi_n(t).$$

In the Bayesian framework, the prior mean $\mu_0(t)$ is our initial guess at the curve. The covariance $\gamma_0(s,t)$ represents the vagueness or uncertainty in our knowledge.

Let $y(t) = x(t) + n(t)$ where $x(t)$ is a Gaussian process with mean $\mu_0(t)$ and covariance $\gamma_0(s,t)$; $n(t)$ is white noise with variance $\sigma^2 > 0$ and independent of $x(t)$. Assume that the observations $y_i$ are discretely sampled points from $y(t)$ at $t_i$, $i = 1, \cdots, N$. The posterior of $x(t)$ given $y_i$, $i = 1, \cdots, N$ is again a Gaussian process (Kimeldorf and Wahba, 1970) with mean

$$\mu_N(t) = \mu_0(t) + \boldsymbol{a}'(t)(\sigma^2 I + K)^{-1}(\boldsymbol{y} - \boldsymbol{\mu}).$$

and covariance

$$\gamma_N(s,t) = \gamma_0(s,t) - \boldsymbol{a}'(s)(\sigma^2 I + K)^{-1}\boldsymbol{a}(t),$$

where $\boldsymbol{y} = (y_1, \cdots, y_N)'$, $\boldsymbol{\mu} = (\mu_0(t_1), \cdots, \mu_0(t_N))'$, $\boldsymbol{a}(t) = (\gamma_0(t,t_1), \cdots, \gamma_0(t,t_N))'$ and $K = (\gamma_0(t_i,t_j))$ is the Gram-Schmidt matrix. This is essentially a Gauss-Markov theorem in the infinite-dimensional parameter situation.

Spline regression also connects to Gaussian process priors. A spline is a piecewise polynomial satisfying certain smooth condition. This concept arises from optimal approximation problems and its applications can be found in many branches of mathematics, statistics and engineering. The Bayesian view of the spline method is made clear by Wahba (1978). If the prior distribution of the regression function is equivalent to that of

$$\sum_{i=0}^{r-1} \theta_i \frac{t^i}{i!} + \frac{\sigma}{\sqrt{N\lambda}} \int_0^t \frac{(t-u)^{r-1}}{(r-1)!} dW(u), \quad \lambda > 0,$$

where $W$ is a Wiener process and $\theta_i \sim N(0,\xi)$ are *iid* and independent of $W$, then the spline estimate of the regression curve with smoothing parameter $\lambda$ is the limit of the Bayesian estimate as $\xi$ tends to infinity. Briefly, the spline regression is a Bayesian estimate under a diffuse or non-informative prior.

## 1.5 The Gibbs Sampler

A current trend in Bayesian computational research has drawn the attention towards the resampling approach, especially the so-called MCMC (Markov Chain Monte Carlo) methods. In fact, the MCMC approach is not new, but only in recent years has it become popular.

The basic idea of the MCMC method is straightforward. Suppose we want to generate a sample from the posterior distribution $\pi$ but cannot do so directly. We proceed by constructing a Markov chain with $\pi$ as its equilibrium distribution. Then the ergodicity of the Markov chain offers an approximation to the posterior distribution $\pi$. For a given $\pi$ there

may be many ways to construct such a Markov chain, and various types of MCMC methods arise from different constructions of the Markov chain. The reader is referred to Smith and Roberts (1993) for a comprehensive review.

One of the MCMC methods which seems especially attractive to Bayesians is the Gibbs sampler, which was originally proposed by Geman and Geman (1984) and later introduced to the statistical literature by Gelfand and Smith (1990). Suppose the joint posterior density for $\theta = (\theta_1, \cdots, \theta_s)$ is $\pi(\theta)$ and the conditional density of $\theta_i$ given $\theta_{[-i]} = (\theta_1, \cdots, \theta_{i-1}, \theta_{i+1}, \cdots, \theta_s)$ is $\pi(\theta_i|\theta_{[-i]})$. The standard Gibbs sampler is an iterative updating scheme described as follows: Initially choose an arbitrary starting value $\theta^{(0)} = (\theta_1^{(0)}, \cdots, \theta_s^{(0)})$ and then update $\theta^{(0)}$ into $\theta^{(1)}$ by generating

$$\begin{aligned}
\theta_1^{(1)} &\sim \pi(\theta_1|\theta_2^{(0)}, \cdots, \theta_s^{(0)}), \\
\theta_2^{(1)} &\sim \pi(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, \cdots, \theta_s^{(0)}), \\
&\cdots \quad \cdots \quad \cdots\cdots\cdots \\
\theta_s^{(1)} &\sim \pi(\theta_s|\theta_1^{(1)}, \cdots, \theta_{s-1}^{(1)}).
\end{aligned}$$

This completes one iteration and the process of updating can be continued. Under mild conditions (Tierney, 1994) the sequence $\theta^{(0)}, \theta^{(1)}, \cdots, \theta^{(n)}, \cdots$ forms a realization of a Markov chain whose equilibrium distribution is $\pi$.

To implement a Gibbs sampler, the conditional distribution $\pi(\theta_i|\theta_{[-i]})$ must be easy to sample. This is not always the case in practice, and discussions are available on facilitating Gibbs sampling. For example, Besag and Green (1993) reviewed and discussed this issue in detail. A very effective way of implementing a Gibbs sampler is via auxiliary variables. In some situations, it is easier to work with the joint density $\pi(\theta, \eta)$ rather than the marginal density $\pi(\theta)$. Here $\eta$ is merely an *auxiliary variable* introduced for convenience. According to Besag and Green (1993), the auxiliary variables help to reduce interaction and thus accelerate the convergence. In some cases, the auxiliary variables also simplify the Gibbs

sampling. Suppose we want to sample

$$\pi(\theta) \propto \frac{\theta_1^{\alpha_1} \cdots \theta_s^{\alpha_s}}{(1+\theta_1+\cdots+\theta_s)^{\beta}}, \quad \theta_i > 0,$$

where $\alpha_i, \beta > 0$. Obviously, the conditional distribution $\pi(\theta_i|\theta_{[-i]})$ is unfamiliar, and thus the original form of Gibbs sampling is not easily carried out. But if we consider the joint density

$$\pi(\theta, \eta) \propto \theta_1^{\alpha_1} \cdots \theta_s^{\alpha_s} \eta^{\beta-1} \exp[-(1+\theta_1+\cdots+\theta_s)\eta], \quad \theta_i > 0, \ \eta > 0$$

with $\pi(\theta)$ as its marginal density, the implementation becomes much easier. The Gibbs sampling between $\theta$ and $\eta$ is automatic and substantially simplified in that only Gamma distributions are involved and updating is between two components only. This technique will be used constantly in this thesis to circumvent difficulties in Gibbs sampling.

## 1.6 Mixture and Hierarchical Models

The class of mixtures of some standard distributions, say the normal or Beta, is very rich in the sense that, there is a member in the class arbitrarily close to any given distribution. For example, the well-known Bernstein theorem states that any continuous function $f$ on the unit interval $[0, 1]$ can be uniformly approximated by a sequence of polynomials, namely, its Bernstein polynomials.

$$B_n(x, f) = \sum_{j=0}^{n} \binom{n}{j} f(\frac{j}{n}) x^j (1-x)^{n-j}, \quad n = 1, 2, \cdots.$$

From the statistical point of view, this can be interpreted as follows: Any probability distribution on the unit interval $[0, 1]$ with continuous density can be uniformly approximated by a finite mixture of Beta distributions.

Escobar (1994) proposed a hierarchical model

$$y_i = \mu_i + \epsilon_i, \quad i = 1, \cdots, n,$$

where $\epsilon_i \sim N(0, \sigma^2)$ and $\sigma^2$ is known. Further, $\mu_i \sim F$ are independent. A Dirichlet process prior is assigned to $F$ representing the initial information. Observations are thus from a mixed normal and prior knowledge is assumed on the mixing distribution rather than directly on the data distribution. Inference on $\mu_i$ is of interest and implemented by the Gibbs sampler. In a more recent work, Escobar and West (1995) extended this study to consider the unknown variance, and examined applications in density estimation.

Mixtures on intervals other than $(-\infty, \infty)$ are formed through Gamma or Beta Distributions. Distributions on $(0, 1)$, or generally a finite interval, can be approximated by a mixture of Beta distributions. For instance, $Beta(n\eta + 1, n(1-\eta) + 1)$ with $\eta$ being a random variable on $(0, 1)$ with density $f$ is close to $f$ when $n$ is large enough.

## 1.7 Statement of the Problem and Outline

It is the purpose of this thesis to develop Bayesian analysis for failure time data and bioassay data. Generally, the bioassay problem still falls in the framework of survival analysis if we follow the concept of tolerance. The approach is basically non-parametric because we do not confine the life distribution or tolerance distribution to any specific parametric class; the advantage is apparent when field knowledge is not enough to determine the type of the life or tolerance distribution. As stated earlier, we would work towards a simple and practical Bayesian analysis for right-censored failure time data using smoothed priors, and for binary and doubly-censored data using Dirichlet process priors.

Due to the complexity of the smoothed prior, some preliminary work is done in Chapter 2. A multivariate Gamma distribution is constructed to incorporate correlation. The related Bessel family which includes the Bessel distribution, squared Bessel process and bridge is introduced. Some numerical computation issues regarding the evaluation of the Bessel quotient and the simulation of Bessel variables are discussed in detail.

Chapter 3 is basically the problem of one-sample survival analysis under a smoothed prior. Posterior computations for right-censored failure time data are efficiently handled and illustrated by numerical examples. Bayesian inference under non-informative priors is addressed and compared with existing results.

Chapter 4 retreats from the infinite-dimensional priors and considers a more practical approach using finite-dimensional priors. Links to some well-known results such as the partial likelihood for proportional hazards regression and the $A(n)$ for prediction are established. Fiducial inference for failure time data is also presented. This is numerically equivalent to Bayesian analysis under a non-informative and data-dependent prior.

Previous research on the application of the Dirichlet process prior in bioassay data is not completely successful, at least in the implementation thereof. The topic of Chapter 5 will be the Bayesian analysis of a single bioassay and the combining of many assays. More complicated data types, such as doubly-censored data, are also considered.

The concluding chapter summarizes the findings in the effort to achieve our goal. Some random thoughts and comments on the current study are presented informally and future research topics are also discussed.

*CHAPTER 2*

# *The Bessel Family and Gamma Distributions*

## 2.1 Introduction

The concept of conjugate priors is central in traditional Bayesian statistics. In nonparametric survival analysis the form of the likelihood varies according to the way we parameterize the model, and the following parameterization suggests using the Gamma distribution as a prior. Suppose we observed failures at times $t_1 < \cdots < t_n$ $(t_0 = 0)$ with $d_i$ subjects failed at $t_i$, and censorings in $[t_i, t_{i+1})$ are adjusted to $t_i$. Let $\Phi_i$ denote the increment of the cumulative hazard over $(t_{i-1}, t_i)$, $1 \leq i \leq n$, and let $\phi_i$ be the hazard at time $t_i$. Then the likelihood for $(\phi, \Phi)$ given the data can be expressed as

$$L(\phi, \Phi) = \phi_1^{d_1} \cdots \phi_n^{d_n} \exp(-s_1\Phi_1 - \cdots - s_n\Phi_n),$$

where $s_i$ denotes the number of subjects at risk just before time $t_i$. Approximately, $\phi_i = \Phi_i/(t_i - t_{i-1})$ and thus the likelihood for $\Phi$ turns out to be

$$L(\Phi) = \Phi_1^{d_1} \cdots \Phi_n^{d_n} \exp(-s_1\Phi_1 - \cdots - s_n\Phi_n),$$

which indicates that an independent Gamma prior is conjugate in this situation.

However, a multivariate Gamma distribution is needed if the relationships between parameters are taken into account. This is the motivation for constructing the exponentially

correlated Gamma distribution in section 2.3. The randomized Gamma distributions arise when we study various conditional distributions. Furthermore, the mechanism of the randomization involves the Bessel distribution, which is relatively unknown, and a general discussion is provided.

The path integral is a device originally created for quantum physics. From the view of probability, it is a conditional Laplace transform. The link between path integrals and some differential equations was exhibited in the 1940's, but the actual evaluation of path integrals is only possible for some special cases. Section 2.5 will show some applications of this device in dealing with some complex distributions.

The numerical evaluation of various Bessel functions has received widespread attention and many articles have been published proposing possible solutions. But the inherent structure of the Bessel functions is so complicated that none of the existing methods is really efficient. It is fortunate that our computations only involve the ratio of two Bessel functions, which is called Bessel quotient, with complexity much less than that of the Bessel function itself. The evaluation of the Bessel quotient is based on its continued fraction representation, a well-known result. In section 2.6 we provide some elementary analytical properties of the Bessel quotient.

Finally, we propose a method for simulating Bessel distributions. The efficiency and accuracy of this Bessel generator is vital in our subsequent posterior computations.

## 2.2 The Bessel Distributions

A random variable $Y$, taking values on the non-negative integers, is said to be a Bessel random variable with parameters $\nu > -1$ and $a > 0$ if

$$\Pr(Y = n) = \frac{1}{I_\nu(a)\, n!\, \Gamma(n + \nu + 1)} \left(\frac{a}{2}\right)^{2n+\nu}, \quad n = 0, 1, \cdots. \tag{2.1}$$

where $I_\nu(x)$ denotes the first type of (modified) Bessel function given by

$$I_\nu(x) = (\frac{x}{2})^\nu \sum_{n=0}^{\infty} \frac{1}{n!\Gamma(n+\nu+1)}(\frac{x}{2})^{2n}, \quad x > 0, \quad \nu > -1.$$

For simplicity we use the notation $Bes(\nu, a)$ for the Bessel distribution with parameters $\nu$ and $a$.

It is obvious that (2.1) gives a probability mass function, but unlike the binomial or Poisson, which arises naturally from some physical process, the definition of the Bessel distribution may seem somewhat artificial. Thus, it is desirable to reveal its many faces and link it to distributions that are familiar to most readers.

(i) *The Bessel distribution as an inverse probability.* Assigning a Gamma prior to the mean of a Poisson distribution is standard in Bayesian statistics; in this example, however, we put a Poisson prior on a Gamma distribution. Suppose we want to draw an inference about the number of customers visiting a laundromat based on the power consumption. The observable total power consumption $Y$ in a period $T$ breaks up into two parts: the customer consumption $Y_1$ and a base amount $Y_2$ independent of $Y_1$. Further, we assume the power consumption of each customer is an exponential random variable with scale $a$, and the distribution of $Y_2$ is $G(\nu+1, a)$ where $\nu > -1$. Let the number of customers $r$ be the parameter of interest. Then, the distribution of $Y$ given $r$ is $G(r+\nu+1, a)$. If customer arrivals are described by a Poisson process with rate $\lambda$, the prior distribution for $r$ should be $P(\lambda T)$. Given an observation $y$, it follows that the posterior distribution of $r$ is $Bes(\nu, 2\sqrt{a\lambda T y})$.

(ii) *The Bessel distribution as a conditional Poisson distribution.* When $\nu$ is an integer, the Bessel distribution $Bes(\nu, a)$ is the conditional distribution of $Y$ given $X - Y = \nu$, where $X \sim P(\lambda_1)$ and $Y \sim P(\lambda_2)$ are independent and $\lambda_1\lambda_2 = a^2/4$.

For the general case $\nu \geq 0$, $X$ is generated from a randomized Poisson distribution. Let $X \sim P(\lambda_1 - \eta)$ with $\eta \sim G(\nu - [\nu], 1)$ but right truncated at $\lambda_1$, where $[\nu]$ denotes the integer part of $\nu$. To include the integer case we adopt the convention that $G(0, 1)$ denotes

the probability distribution concentrated on zero. Now, the density of $\eta$ is proportional to $\eta^{\nu-[\nu]-1}e^{-\eta}I(0<\eta<\lambda_1)$, so that

$$\begin{aligned}\Pr(X=k) &\propto \int \Pr(X=k|\eta)\eta^{\nu-[\nu]-1}e^{-\eta}I(0<\eta<\lambda_1)d\eta \\ &= \frac{e^{-\lambda_1}}{\Gamma(k+1)}\int_0^{\lambda_1}\eta^{\nu-[\nu]-1}(\lambda_1-\eta)^k d\eta \\ &\propto \frac{\lambda_1^{k+\nu-[\nu]}}{\Gamma(k+1+\nu-[\nu])}.\end{aligned}$$

It is now clear that, for $\nu \geq 0$, the conditional distribution of $Y$ given $X - Y = [\nu]$ is $Bes(\nu, a)$. We now turn to build a kind of recurrence relation for Bessel distributions.

(iii) *The Bessel distribution as a sum of Bernoulli variables.* It is well-known that the Bessel function satisfies the recurrence equation

$$I_\nu(x) = I_{\nu+2}(x) + \frac{2(\nu+1)}{x}I_{\nu+1}(x). \tag{2.2}$$

which implies a kind of relation between $Bes(\nu, a)$, $Bes(\nu+1, a)$ and $Bes(\nu+2, a)$. In fact, it is immediately seen that the Bessel distribution $Bes(\nu, a)$ is a mixture of $Bes(\nu+1, a)$ and a right-shifted $Bes(\nu+2, a)$ produced by moving the mass at $k$ to $k+1$. The two weights for this mixture are $2(\nu+1)R_\nu(a)/a$ and $R_\nu(a)R_{\nu+1}(a)$ respectively, where $R_\nu(a) = I_{\nu+1}(a)/I_\nu(a)$ is called the Bessel quotient. In the language of sampling, a random variable $Y \sim Bes(\nu, a)$ can be generated by first generating a Bernoulli random variable $r$ with parameter $\Pr(r = 1) = R_\nu(a)R_{\nu+1}(a)$ followed by $X \sim Bes(\nu + r + 1, a)$, and then $Y = X + r$.

From this property, a Bessel random variable can be expressed as a sum of Bernoulli variables: First, a random variable $Y \sim Bes(\nu, a)$ can be written as $Y = r_1 + X_1$ with $r_1$ a Bernoulli variable with parameter $R_\nu(a)R_{\nu+1}(a)$ and $X_1 \sim Bes(\nu + r_1 + 1, a)$. Then, $X_1$ can be written as $X_1 = r_2 + X_2$ with $r_2$ a Bernoulli variable with parameter $R_{\nu+r_1+1}(a)R_{\nu+r_1+2}(a)$ and $X_2 \sim Bes(\nu + r_1 + r_2 + 2, a)$. Since $Bes(\nu + k, a)$ can be treated as a point mass on zero for $k$ large enough, we can express $Y$ as an infinite sum of Bernoulli variables $\sum_{i=1}^{\infty} r_i$.

(iv) *Relationship to the von Mises distribution.* The von Mises distribution, which was introduced by von Mises, is an analogue of the normal distribution in circular statistics. Its

density function is

$$\phi(\theta) = \frac{1}{2\pi I_0(\kappa)} \exp(\kappa \cos\theta), \quad -\pi \le \theta < \pi, \ \kappa > 0,$$

where $\kappa$ is the concentration parameter. Detailed study and interesting applications can be found in Mardia (1972). It is not surprising that the Bessel distribution offers a simple characterization for the von Mises distribution. If $\theta$ is a von Mises variable with concentration parameter $\kappa$, then the distribution of $\cos^2\theta$ is a randomized Beta distribution $Beta(r + 1/2, 1/2)$ where $r \sim Bes(0, \kappa)$. However, given $\cos^2\theta = y$ there are still four possible values of $\theta$ in $[-\pi, \pi)$. The uncertainty can be removed given the sign of $\cos\theta$ and the sign of $\theta$ itself. To determine the sign of $\theta$ we can simply flip a fair coin since the von Mises distribution is symmetric. But to determine the sign of $\cos\theta$ we need a biased coin with head and tail probabilities proportional to $(e^{\kappa\sqrt{y}}, e^{-\kappa\sqrt{y}})$ which is based on the density of the von Mises distribution. Now, suppose $Y \sim Beta(r+1/2, 1/2)$ where $r \sim Bes(0, \kappa)$, and given $Y$, $b_1$ and $b_2$ are independent Bernoulli variables with parameters $1/2$ and $1/(1+e^{-2\kappa\sqrt{Y}})$ respectively. Then,

$$\theta = (2b_1 - 1)\arccos[(2b_2 - 1)\sqrt{Y}]$$

is a von Mises variable with concentration parameter $\kappa$.

(v) *Moments and mode.* The moments of the Bessel distribution can be expressed in terms of the Bessel quotient. For instance, if $Y \sim Bes(\nu, a)$, then

$$E\,Y = \frac{1}{2} a R_\nu(a) \quad \text{and} \quad E\,Y^2 = \frac{1}{4} a^2 R_\nu(a) R_{\nu+1}(a) + \frac{1}{2} a R_\nu(a). \tag{2.3}$$

The factorial moments,

$$E\,Y(Y-1)\cdots(Y-k+1) = (\frac{a}{2})^k \, R_\nu(a)\cdots R_{\nu+k-1}(a), \quad k = 1, 2, \cdots,$$

are easily obtained and from these we can calculate the moment of any order.

Finally, the Bessel distribution has a unique mode, or two modes at consecutive integers. For convenience we make the convention that the mode of a Bessel distribution always refers

to the larger one if there are two modes and, it then follows that the mode of $Bes(\nu, a)$ is the integer part of $m(\nu, a) = (\sqrt{a^2 + \nu^2} - \nu)/2$. This is useful in simulating the Bessel distribution.

## 2.3 Multivariate and Randomized Gamma Distributions

We begin with a simple construction of a bivariate Gamma distribution. Consider the joint Laplace transform of independent random variables $Y_1 \sim G(\alpha, \lambda_1)$ and $Y_2 \sim G(\alpha, \lambda_2)$.

$$E\, e^{-t_1 Y_1 - t_2 Y_2} = (1 + \frac{t_1}{\lambda_1})^{-\alpha}(1 + \frac{t_2}{\lambda_2})^{-\alpha} = [\det(I_2 + AT)]^{-\alpha}. \tag{2.4}$$

where $I_2$ is a $2 \times 2$ identity matrix. $A$ and $T$ are $2 \times 2$ diagonal matrixes with entries $1, \lambda_1, 1, \lambda_2$ and $t_1, t_2$ respectively. The simple form of the joint Laplace transform follows from the independence of $Y_1$ and $Y_2$. Our purpose is to construct a bivariate Gamma distribution that accommodates fairly general dependence, and there are many ways to accomplish this. One simple way, however, is to alter the matrix $A$ to a positive symmetric matrix. Therefore, a more general form of $A$ can be obtained if we replace the two zeroes in $A$ by $\sqrt{\rho\lambda_1\lambda_2}$ with $0 \le \rho < 1$. The corresponding Laplace transform is

$$[1 + \frac{t_1}{\lambda_1} + \frac{t_2}{\lambda_2} + \frac{(1-\rho)t_1 t_2}{\lambda_1 \lambda_2}]^{-\alpha}. \tag{2.5}$$

We can find the inverse of (2.5) in two steps. First treat $t_2$ as a constant and calculate the inverse with respect to $t_1$, and then invert the result with respect to $t_2$. Lengthy but straightforward calculation shows that the density function corresponding to the Laplace transform (2.5) is proportional to

$$(y_1 y_2)^{(\alpha-1)/2} \exp(-\frac{\lambda_1 y_1 + \lambda_2 y_2}{1-\rho})\, I_{\alpha-1}(\sqrt{\frac{4\rho\lambda_1\lambda_2 y_1 y_2}{1-\rho}}), \quad y_1, y_2 > 0. \tag{2.6}$$

The density (2.6) keeps the marginal distributions of $Y_1$ and $Y_2$ and, the only thing to be investigated is the dependence imposed by this kind of generalization. Specifically, we

are interested in the conditional distribution of $Y_1$ given $Y_2$, or vice versa. However, the seemingly complicated form of the conditional density could obscure a simple fact if we do not go a step further. For this purpose, the notion of the randomized Gamma distribution (Feller, 1966) turns out to be helpful.

Suppose that $Y|r \sim G(\alpha + r, \lambda)$ with $\alpha, \lambda > 0$ and $r \sim P(a/4\lambda)$ where $a > 0$. Then the marginal distribution of $Y$ is called a randomized Gamma distribution of the first type. One can easily verify that the density of the randomized Gamma distribution is proportional to $y^{(\alpha-1)/2}e^{-\lambda y}I_{\alpha-1}(\sqrt{ay})$, $y > 0$, a functional form we shall meet many times.

It now becomes clear that the conditional distribution of $Y_1$ given $Y_2 = y_2$ is a randomized Gamma distribution $G(\alpha + r, \lambda_1/(1-\rho))$ with $r \sim P(\rho\lambda_2 y_2)$. The constant $\rho$ measures the degree of dependence between $Y_1$ and $Y_2$ and is thus called the dependence coefficient. A larger value of $\rho$ indicates a stronger dependence and $\rho = 0$ corresponds to independence.

Can we construct a multivariate Gamma distribution in this way? Conceptually, there is no difficulty for the matrixes in (2.4) can be of any size. The problem is that a closed form density may not always be available. We restrict our attention to a special case, namely, the multivariate Gamma with Markov dependence where a closed form density does exist.

A random vector $(Y_1, \cdots, Y_n)$ is said to be Markov if the conditional distribution of $Y_i$ given $Y_1, \cdots, Y_{i-1}$ is the same as that of $Y_i$ given $Y_{i-1}$. Let $\rho_i$ denote the dependence coefficient between $Y_{i-1}$ and $Y_i$. If we wish to ensure that each marginal distribution of $Y_i$ is $G(\alpha, \lambda_i)$, then under the the assumption of Markov dependence, the joint density of $(Y_1, \cdots, Y_n)$ is proportional to

$$(y_1 y_n)^{(\alpha-1)/2} \exp[-\sum_{i=1}^{n} \frac{(1-\rho_i\rho_{i+1})\lambda_i y_i}{(1-\rho_i)(1-\rho_{i+1})}] \prod_{i=1}^{n-1} I_{\alpha-1}(\sqrt{\frac{4\rho_{i+1}\lambda_i\lambda_{i+1}y_i y_{i+1}}{1-\rho_{i+1}}}), \quad y_i > 0. \quad (2.7)$$

where $\rho_1 = \rho_{n+1} = 0$.

The randomized Gamma distribution of the second type is the mixture distribution $G(\alpha + r_1 + 2r_2, \lambda)$, $\alpha, \lambda > 0$, where $r_1$ and $r_2$ are independent with Poisson and Bessel distributions

respectively. This can be viewed as a generalization of the first type taking $r_2$, in a limiting case, as zero. For any positive numbers $a$, $b$, $\lambda$ and $\alpha$, the randomized Gamma $G(\alpha + r_1 - 2r_2, \lambda)$ with $r_1 \sim P((a+b)/(4\lambda))$ and $r_2 \sim Bes(\alpha - 1, \sqrt{ab}/(2\lambda))$ independent has a density function proportional to $e^{-\lambda y} I_{\alpha-1}(\sqrt{ay}) I_{\alpha-1}(\sqrt{by})$, $y > 0$. A proof of this will be provided later in section 2.4.

A randomized Gamma distribution of the second type arises from (2.7) when we consider the conditional distribution of $Y_i$ given $Y_{i-1} = y_{i-1}$ and $Y_{i+1} = y_{i+1}$. The conditional density is proportional to

$$\exp[-\frac{(1-\rho_i\rho_{i+1})\lambda_i y_i}{(1-\rho_i)(1-\rho_{i+1})}]\, I_{\alpha-1}(\sqrt{\frac{4\rho_i\lambda_{i-1}\lambda_i y_{i-1} y_i}{1-\rho_i}})\, I_{\alpha-1}(\sqrt{\frac{4\rho_{i+1}\lambda_i\lambda_{i+1} y_i y_{i+1}}{1-\rho_{i+1}}}), \quad y_i > 0,$$

exactly the form given above.

## 2.4 The Squared Bessel Processes

For any $d > 0$, the $d$-dimensional squared Bessel process $\xi(t)$, $t \geq 0$ is a time homogeneous Markov process with transition density

$$q(t, x, y) = \frac{1}{2t}(\frac{y}{x})^{\nu/2} \exp(-\frac{x+y}{2t}) I_\nu(\frac{\sqrt{xy}}{t}), \quad t > 0, \ x, y \geq 0, \tag{2.8}$$

where $\nu + 1 = d/2$. From the last section, the conditional distribution of $\xi(t)$ given $\xi(0) = x$ is a randomized Gamma distribution of the first type $G(\nu + r + 1, 1/2t)$ with $r \sim P(x/2t)$. When $d$ is an integer and $\xi(0) = 0$, $\xi(t)$ can be expressed as

$$\xi(t) = B_1^2(t) + \cdots + B_d^2(t), \quad t \geq 0,$$

where $(B_1(t), \cdots, B_d(t))$, $t \geq 0$ is a standard $d$-dimensional Brownian motion. The sample path of the squared Bessel process is continuous but nowhere differentiable.

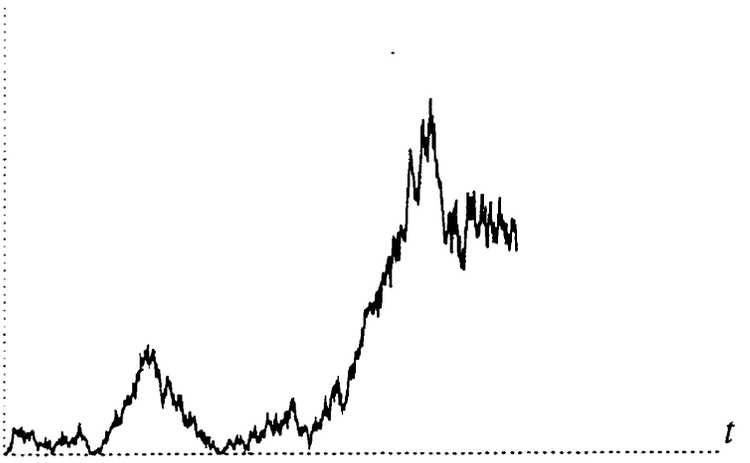


*Fig.* 2.1 A sample path of the squared Bessel process when $d = 1$.

It is not trivial to verify that (2.8) is a transition density. Assuming it is, however, we can derive the density function for the randomized Gamma of the second type. Applying the Chapman-Kolmogorov equation to (2.8) we have

$$\int q(t, x, y)q(t, y, z)dy = q(2t, x, z).$$

Letting $a = x/t^2$, $b = z/t^2$ and $\lambda = 1/t$, we see that

$$\int e^{-\lambda y} I_\nu(\sqrt{ay}) I_\nu(\sqrt{by})dy = \frac{1}{\lambda} \exp(\frac{a+b}{4\lambda}) I_\nu(\frac{\sqrt{ab}}{2\lambda}). \tag{2.9}$$

thus,

$$g(y|a, b, \lambda) = \lambda \exp(-\frac{a+b}{4\lambda}) I_\nu^{-1}(\frac{\sqrt{ab}}{2\lambda}) e^{-\lambda y} I_\nu(\sqrt{ay}) I_\nu(\sqrt{by})$$

is a probability density function.

Using (2.9) again we can calculate the Laplace transform

$$\int e^{-ty} g(y|a, b, \lambda)dy = \frac{\lambda}{\lambda + t} \exp[\frac{(a+b)}{4\lambda}(\frac{\lambda}{\lambda + t} - 1)] I_\nu(\frac{\sqrt{ab}}{2(\lambda + t)}) I_\nu^{-1}(\frac{\sqrt{ab}}{2\lambda}).$$

This can be seen to be identical to the randomized Gamma $G(\alpha + r_1 + 2r_2, \lambda)$ with $r_1 \sim P((a+b)/(4\lambda))$ and $r_2 \sim Bes(\alpha - 1, \sqrt{ab}/(2\lambda))$ independent. Thus $g$ is the density of the randomized Gamma distribution described above.

When $\xi(0) = 0$ the marginal distribution of $\xi(t)$ is $G(d/2, 1/(2t))$ and, for $s < t$ the joint distribution of $\xi(s)$ and $\xi(t)$ has a density with the same form as (2.6) with the parameters given by $\alpha = d/2$, $\lambda_1 = 1/(2s)$, $\lambda_2 = 1/(2t)$ and $\rho = s/t$. There is a correspondence between a multivariate Gamma given by (2.7) and a re-scaled finite dimensional distribution of a Bessel process. Specifically, if we sample $\xi(t)$ at times $t_i = 1/(\rho_1 \cdots \rho_i)$, $i = 1, \cdots, n$, then the joint distribution of $\xi(t_1)/(t_1\lambda_1), \cdots, \xi(t_n)/(t_n\lambda_n)$ is exactly the same as (2.7) provided $d = 2\alpha$ and $\lambda_i > 0$, $0 < \rho_i < 1$.

A standard squared Bessel bridge $\xi_{x_0,x_1}(t)$ is a stochastic process on $[0, 1]$ generated by $\xi(t)$ with $\xi(0)$ and $\xi(1)$ tied at $x_0$ and $x_1$ respectively. The distribution of $\xi_{x_0,x_1}(t)$ is studied in detail by Pitman and Yor (1982). Its transition distribution is a randomized Gamma distribution of the second type. Let $0 \leq s < t \leq 1$. If we know that $\xi_{x_0,x_1}(s) = x$, then, $Y = \xi_{x_0,x_1}(t)$ can be obtained by generating independent random variables

$$r_1 \sim P(\frac{1}{2(1-s)}[\frac{(1-t)}{(t-s)}x + \frac{(t-s)}{(1-t)}x_1]) \quad \text{and} \quad r_2 \sim Bes(\nu, \frac{\sqrt{xx_1}}{1-s})$$

and then

$$Y \sim G(\nu + r_1 + 2r_2 + 1, \frac{1-s}{2(t-s)(1-t)}).$$

## 2.5 Path Integrals

The path integral or Feynman integral (Gel'fand and Yaglom, 1960) is a device used in quantum physics for integration over a function space. The first result of path integral evaluation (Cameron and Martin, 1944) is the so-called Cameron-Martin formula:

$$E \exp[-\lambda \int_0^1 B^2(t)dt] = (\cosh \sqrt{2\lambda})^{-1/2}.$$

where $B(t)$ is a standard Brownian motion and $\lambda > 0$. The path integrals for the Bessel process can be found in Pitman and Yor (1982). Actually, for fixed initial state the path integral has been considered in a more general setting. It has been shown (Pitman and Yor, 1982) that, for a Radon measure $\mu$ with support in $[0, a]$, the Schrödinger equation

$$(-\frac{1}{2}\Delta + \mu)u = 0. \tag{2.10}$$

$$u(0) = 1, \quad u'(a) = 0 \tag{2.11}$$

has unique solution $u$ in the sense of distributions on the space $C_0^\infty(0, \infty)$, and

$$E_x \exp[-\int_0^a \xi(t)d\mu(t)] = u^{d/2}(a)\exp[\frac{1}{2}u'(0)x]. \tag{2.12}$$

An immediate result is a generalization of the Cameron-Martin formula to the squared Bessel process:

$$E \exp[-\lambda \int_0^t \xi(s)ds] = (\cosh\sqrt{2\lambda}\, t)^{-d/2}, \tag{2.13}$$

where $\lambda, d > 0$. We will see that including a weighting measure $\mu$ makes (2.12) more powerful than the classical Cameron-Martin formula. For fixed initial and end states, Pitman and Yor (1982) provided a very useful path integral of the squared Bessel process:

$$E_x\,(\exp[-\lambda \int_0^t \xi(s)ds] \mid \xi(t) = y) = \psi_\lambda(t, x, y)$$

where $\lambda > 0$ and

$$\psi_\lambda(t, x, y) = \frac{\sqrt{2\lambda}\, t}{\sinh\sqrt{2\lambda}\, t}\exp\{\frac{x+y}{2t}(1 - \sqrt{2\lambda}\, t\coth\sqrt{2\lambda}\, t)\} I_\nu(\frac{\sqrt{2\lambda xy}}{\sinh\sqrt{2\lambda}\, t}) I_\nu^{-1}(\frac{\sqrt{xy}}{t}). \tag{2.14}$$

The conditional distribution of $\int_0^t \xi(s)ds$ given $\xi(0) = x$ and $\xi(t) = y$ is also available since the path integral is the corresponding Laplace transform. Using the infinite product (Titchmarsh, 1939) formula

$$\frac{\sinh z}{z} = \prod_{k=1}^{\infty}(1 + \frac{z^2}{k^2\pi^2}). \tag{2.15}$$

we see that $\sqrt{2\lambda}t/\sinh\sqrt{2\lambda}t$, the first factor in (2.14), is the Laplace transform of an infinite convolution of $G(1, k^2\pi^2/(2t))$, $k = 1, 2, \cdots$. Next, using the following series expansion (Titchmarsh, 1939)

$$z\coth z = 1 + \sum_{k=1}^{\infty} \frac{2z^2}{z^2 + k^2\pi^2}. \tag{2.16}$$

we can show that the exponential factor in (2.14) is the Laplace transform of an infinite convolution of randomized Gamma distributions $G(r_{1,k}, k^2\pi^2/(2t))$, $k = 1, 2, \cdots$, with $r_{1,k} \sim P((x+y)/t)$, $k \geq 1$ independent. Similarly, the remaining part in (2.14) is the Laplace transform of the convolution of the randomized Gamma distributions $G(\nu + 2r_2, k^2\pi^2/(2t))$, $k = 1, 2, \cdots$, with $r_2 \sim Bes(\nu, \sqrt{xy}/t)$. Therefore, a sample $Y$ from the conditional distribution of $\int_0^t \xi(s)ds$ given $\xi(0) = x$ and $\xi(t) = y$ can be drawn by the following procedure:

(i) *generate an iid random sequence* $r_{1,k} \sim P((x+y)/t)$, $k \geq 1$;

(ii) *generate a random variable* $r_2 \sim Bes(\nu, \sqrt{xy}/t)$ *independent of* $r_{1,k}$, $k \geq 1$;

(iii) *generate an independent random sequence* $\eta_k \sim G(r_{1,k} + 2r_2 + \nu + 1, \pi^2/(2t^2))$, $k \geq 1$;

(iv) $Y = \sum_{k=1}^{\infty} \eta_k/k^2$.

Another path integral

$$U_\lambda(x, y) = E_x\left(\exp[-\lambda\int_0^1 \frac{\xi(t)}{(t+p)^2}dt] \mid \xi(1) = y\right)$$

is related to the posterior calculation in the next chapter. This can be evaluated using (2.12) (see Appendix), yielding the result

$$\begin{aligned} U_\lambda(x, y) &= \frac{\sqrt{8\lambda+1}\sinh\gamma}{\sinh\sqrt{8\lambda+1}\,\gamma}\exp[\frac{xe^{\gamma} + ye^{-\gamma}}{2}\sinh\gamma(\coth\gamma \\ &- \sqrt{8\lambda+1}\coth\sqrt{8\lambda+1}\,\gamma)]I_\nu(\frac{\sqrt{xy(8\lambda+1)}\sinh\gamma}{\sinh\sqrt{8\lambda+1}\,\gamma})I_\nu^{-1}(\sqrt{xy}). \end{aligned} \tag{2.17}$$

where $e^{2\gamma} = (p+1)/p$.

Similarly, we can describe the corresponding conditional distribution by the Laplace transform (2.17), infinite product (2.15) and series (2.16). A sample $Y$ from the conditional distribution of $\int_0^1 \xi(t)(t+p)^{-2}dt$ ($e^{2\gamma} = (p+1)/p$) given $\xi(0) = x$ and $\xi(1) = y$ can be drawn by generating

(i) *an independent random sequence* $\tau_{1,k}$, $k \geq 1$ *where* $\tau_{1,k}$ *has a Poisson distribution with mean*

$$\frac{k^2\pi^2}{k^2\pi^2+\gamma^2}\frac{\sinh\gamma}{\gamma}(xe^{\gamma}+ye^{-\gamma});$$

(ii) $\tau_2 \sim Bes(\nu, \sqrt{xy})$ *independent of* $\tau_{1,k}$, $k \geq 1$;

(iii) *an independent random sequence* $\eta_k \sim G(\tau_{1,k}+2\tau_2+\nu+1, 1/(8\gamma^2))$, $k \geq 1$;

(iv) $Y = \sum_{k=1}^{\infty} \eta_k/(k^2\pi^2+\gamma^2)$.

## 2.6 The Bessel Quotient

Evaluation of the Bessel functions is avoided here. This is primarily due to the inefficiency of existing numerical methods. Evaluating a Bessel function once or twice costs very little using a computer and a mathematical package. But if a resampling scheme requires evaluation of many Bessel functions at every iteration, it could be disastrous. Fortunately, we only need a ratio of two Bessel functions, the Bessel quotient $R_\nu(x) = I_{\nu+1}(x)/I_\nu(x)$, which is substantially easier to work with.

We present some results that are neither trivial nor readily available in the literature. The properties stated here of the Bessel quotient are, however, entirely based on the existing theory.

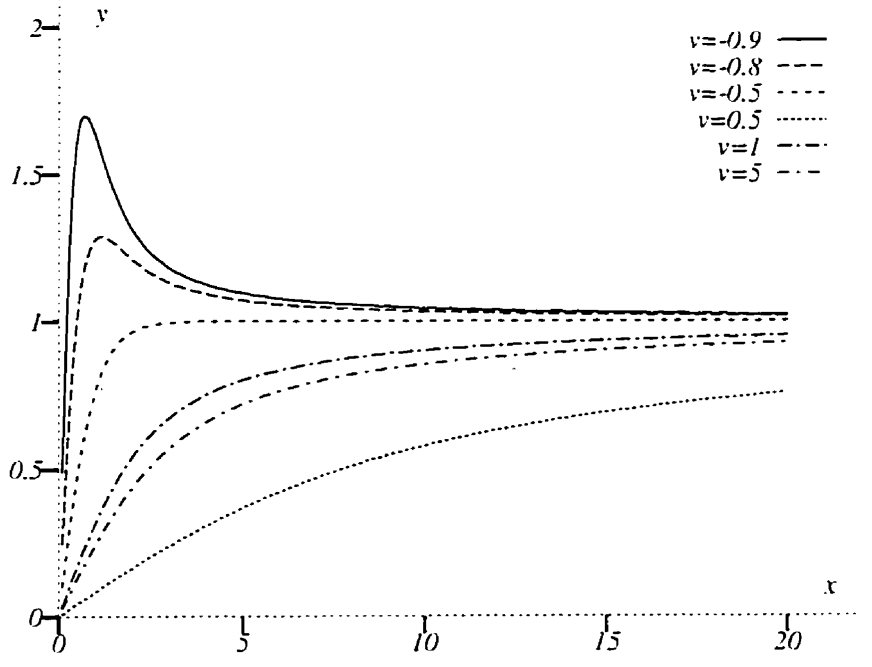


*Fig.* 2.2 The curves of the Bessel quotients for different $\nu$ values.

A recurrence formula for Bessel quotients arises immediately from (2.2):

$$R_\nu(x) = \frac{1}{\frac{2(\nu+1)}{x} + R_{\nu+1}(x)}. \tag{2.18}$$

Further, the following relation for Bessel functions

$$\frac{I'_\nu(x)}{I_\nu(x)} = \frac{\nu}{x} + R_\nu(x), \quad \nu > -1$$

leads to a differential equation

$$\begin{aligned} y' &= 1 - y^2 - \frac{2\nu+1}{x}y. \\ y(0) &= 0 \end{aligned} \tag{2.19}$$

which has unique solution $R_\nu(x)$. These equations are generally important in studying the Bessel quotient. On the other hand, some special Bessel quotients do have closed form

expressions. For example, the integral representation of the Bessel function (Spain and Smith, 1970) states, for $\nu > -\frac{1}{2}$, that

$$I_\nu(x) = \frac{1}{\sqrt{\pi}\Gamma(\nu+\frac{1}{2})}(\frac{x}{2})^\nu \int_{-1}^{1} e^{-xt}(1-t^2)^{\nu-\frac{1}{2}}dt$$

which gives

$$I_{\frac{1}{2}}(x) = \sqrt{\frac{2}{\pi x}}\sinh x, \quad \text{and} \quad I_{\frac{3}{2}}(x) = \sqrt{\frac{2}{\pi x^3}}(x\cosh x - \sinh x).$$

It follows that $R_{\frac{1}{2}}(x) = \coth x - 1/x$ and, by (2.18), $R_{-\frac{1}{2}}(x) = \tanh x$.

The asymptotic expansion (Spain and Smith, 1970) of the Bessel function implies that $R_\nu(x) \to 1$ as $x \to \infty$, and thus all curves of functions $R_\nu(x)$, $\nu > -1$ start from $(0,0)$ and share the same asymptote $y = 1$. Further, we have

$$\lim_{x\to\infty} x[1 - R_\nu(x)] = 2\nu + 1 \tag{2.20}$$

which means the curve of $R_\nu(x)$ approaches the asymptote from above when $-1 < \nu < -\frac{1}{2}$ and from below when $\nu > -\frac{1}{2}$.

The monotonicity is also classified into two categories according to whether $\nu$ is larger than $-\frac{1}{2}$ or not. For $\nu > -\frac{1}{2}$, the function $R_\nu(x)$ is increasing over the whole interval $(0,\infty)$; while for $-1 < \nu < -\frac{1}{2}$, the function $R_\nu(x)$ is increasing first to reach a maximum and then decreasing.

To verify this analytically we differentiate (2.19) to obtain

$$y'' = \frac{2\nu+1}{x^2}y - (2y + \frac{2\nu+1}{x})\,y'.$$

The sign of the second derivative at stationary points, where the first derivative vanishes, is the same as that of $2\nu + 1$ and, this gives rise to an essential difference between $R_\nu(x)$ when $\nu > -\frac{1}{2}$ and $\nu < -\frac{1}{2}$, namely that the former can have local minima only while the latter local maxima only.

From the differential equation (2.19), $y'$ cannot change sign when $\nu > -\frac{1}{2}$. Otherwise, there must be a stationary point $x_0$ which is, according to the discussion above, a local minimum where $y'$ changes from negative to positive. Since $y'(0) = 1/(2\nu + 2) > 0$, there must be a local maximum between zero and $x_0$ which is impossible.

When $-1 < \nu < -\frac{1}{2}$, (2.20) indicates that $y$ must be larger than one when $x$ is large enough and $y(\infty) = 1$. Hence, there must be a point at which $y'$ is negative. Since $y'(0) > 0$, there must be a stationary point $x_0$ which is a local maximum. Furthermore, $y(x_0) > 1$ is guaranteed by (2.19). We can show that $y'$ changes sign only once, otherwise we will have two local maxima between which there must be a local minimum, which is a contradiction. Hence, $y(x_0)$ is also a global maximum.

We are now ready to give some bounds for the Bessel quotient. In fact, the variance of the Bessel distribution calculated from (2.3) must be non-negative and this implies

$$R_\nu(x) \leq \frac{x}{\sqrt{x^2 + \nu^2 + \nu}}. \tag{2.21}$$

On the other hand, the differential equation (2.19) suggests that $R_\nu(x) \leq R_\mu(x)$ if $\nu > \mu$ for these two functions have the same initial value but the former has a smaller derivative. Hence, $R_{\nu+1}(x) \leq R_\nu(x)$ and this, combined with (2.18) and (2.21), leads to

$$\frac{x}{\sqrt{x^2 + (\nu+1)^2} + (\nu+1)} \leq R_\nu(x) \leq \frac{x}{\sqrt{x^2 + \nu^2 + \nu}}, \quad \nu > -1, \tag{2.22}$$

which gives upper and lower bounds. For $\nu > -\frac{1}{2}$ a slightly sharper upper bound can be derived from the fact $R'_\nu(x) \geq 0$. This kind of bounds was also derived by Amos (1974) for the case $\nu \geq 0$.

From (2.18) we see that $R_\nu(x)$ can be computed if we know the value of $R_{\nu+k}(x)$ for some $k$. Actually, by repeating the recurrence relation, $R_\nu(x)$ can be written as a continued

fraction

$$R_\nu(x) = \cfrac{1}{\cfrac{2(\nu+1)}{x} + \cfrac{1}{\cfrac{2(\nu+2)}{x} + \cfrac{1}{\cfrac{2(\nu+3)}{x} + \cdots}}}$$

or, in a compact notation,

$$R_\nu(x) = \frac{1}{2(\nu+1)/x+} \quad \frac{1}{2(\nu+2)/x+} \quad \frac{1}{2(\nu+3)/x+\cdots}. \tag{2.23}$$

The upper bound in (2.22) implies that, for fixed $x$, $R_{\nu+k}(x) \to 0$ as $k \to \infty$ and it seems that $R_\nu(x)$ can be computed by iteration.

The validity of (2.23), or in other words, the convergence of the continued fraction is easily verified. Practically, however, (2.23) is not directly applicable for the round-off error by iteration could pose a problem. A coupled iteration (Amos, 1974) is more appropriate.

The continued fraction is a powerful device for numerical computation. Elementary functions such as $e^x$ and $\tan x$ are actually evaluated based on their continued fractions rather than Taylor series. We hope this short discussion has given a clear picture of the Bessel quotient, a function we will meet again.

## 2.7 Simulating Bessel Distributions

Simulating a Bessel Distribution is generally difficult due to its complexity and loose links with well-known distributions. Standard procedures such as the discrete inverse integral transform, or rejection sampling can be used to generate Bessel random variables. But none of these is easily implemented. The method proposed here for simulating Bessel Distributions is a compromise between efficiency and accuracy.

It seems that a normal approximation is applicable to a fairly large class of Bessel dis-

tributions. Specifically, we use the normal distribution left-truncated at zero with density

$$f(x|\mu,\sigma^2) = \begin{cases} \dfrac{\phi((x-\mu)/\sigma)}{\sigma[1-\Phi(-\mu/\sigma)]} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

where $\Phi$ and $\phi$ are the cumulative distribution function and density function of the standard normal respectively. To avoid the trouble of solving some complicated equations, the parameters $\mu$ and $\sigma^2$ are simply set to be the mean and variance of the Bessel given by (2.3). One must understand that this choice of parameters is not strictly optimal since $\mu$ and $\sigma^2$ are not the mean and variance of the truncated normal. But in practice the inaccurate approximation will not be used. The only loss is that a Bessel distribution which is deemed to lack a good normal approximation might actually have one if the parameters were chosen more carefully.

An intuitive argument for using the normal approximation is available from section 2.2. Note that, for integer-valued $\nu$ the Bessel distribution $Bes(\nu, a)$ is the conditional distribution of $Y$ given $X - Y = \nu$, where $X \sim P(a/2)$ and $Y \sim P(a/2)$ are independent random variables. It is well-known that the Poisson distribution with large mean is similar to a normal distribution. Hence, the $Bes(\nu, a)$ should be similar to a normal distribution when $a$ is large. For example, $Bes(0, 12)$ displayed in Figure 2.3 (I) is very close to a normal distribution. But generally the normal approximation to a Poisson distribution is only accurate in a region centered at its mean, and less accurate in the two tails. A large value of $a$ is not the sole condition for a satisfactory normal approximation. Intensive numerical experiments show that the normal approximation is accurate only when $a$ is large and $\nu$ is relatively small. In the Bessel distribution $Bes(\nu, a)$, increasing $a$ without changing $\nu$ will improve the normal approximation; on the other hand, increasing $\nu$ for fixed $a$ will decrease its accuracy.

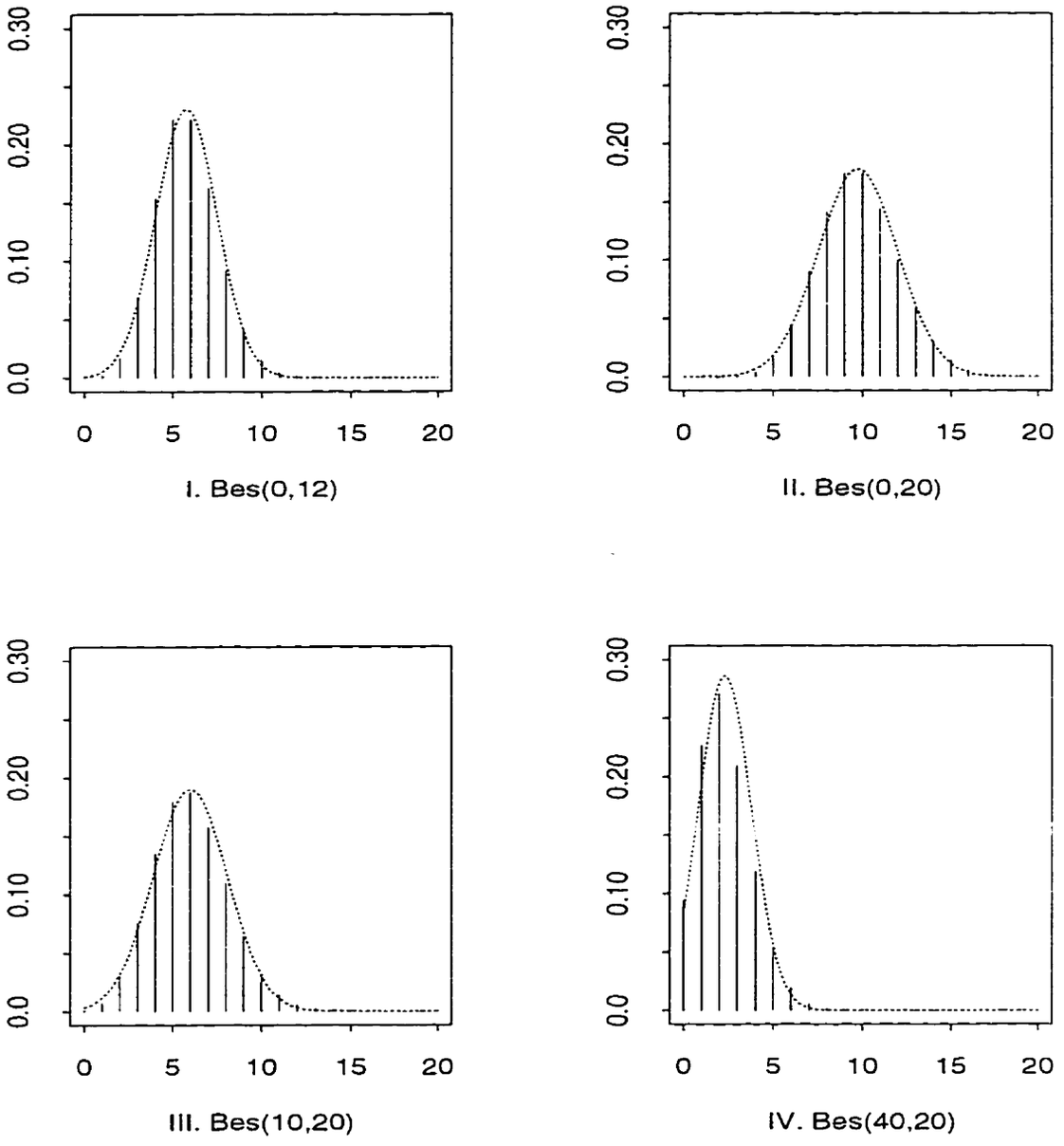


*Fig.* 2.3 The Bessel distributions and their normal approximations.

For example, as shown by Figure 2.3 (II) and (III), $Bes(0, 20)$ is more similar to a normal than $Bes(0, 12)$ and, $Bes(10, 20)$ is less similar to normal than $Bes(0, 20)$. If we increase the value of $\nu$ to 40 we can see that the normal approximation in Figure 2.3 (IV) is obviously not satisfactory.

For any fixed $\nu$ there is a threshold value $a(\nu)$ such that the normal approximation is acceptable for all $Bes(\nu, a)$, $a \geq a(\nu)$. The Bessel distributions are thus divided into two groups, one that is close to the approximating normal distribution and one that is not. If we can find a proper threshold $a(\nu)$ we can then officially define these two groups.

Note that the mode of $Bes(\nu, a)$, or the quantity $m(\nu, a)$, can be written as

$$m(\nu, a) = \frac{a}{2} \frac{a}{\sqrt{a^2 + \nu^2} + \nu}.$$

The first factor measures the magnitude of $a$ and, the second factor, which is between zero and one, measures the relative magnitude of $\nu$ to $a$. A Bessel distribution $Bes(\nu, a)$ with a large mode has to meet both requirements: a large $a$ and a relatively small $\nu$. Hence, the mode might set a threshold but how large it has to be remains a problem. Again, we find the answer from numerical experiment that the mode should be no less than 6 or, equivalently, $m(\nu, a) \geq 6$. Hence, the threshold should be $a(\nu) = \sqrt{24(\nu + 6)}$.

When $m(\nu, a) \geq 6$, a sample $Y \sim Bes(\nu, a)$ is drawn by generating $U \sim U(0, 1)$ followed by

$$X = \mu + \sigma \Phi^{-1}[U + (1 - U)\Phi(-\mu/\sigma)]$$

where $\mu = \frac{1}{2} a R_\nu(a)$ and $\sigma^2 = \frac{1}{4} a^2 - \mu^2 - \nu\mu$; $Y$ is then identified as the closest integer to $X$. This can be easily implemented since $\Phi$ and its inverse $\Phi^{-1}$ are available in major statistical packages.

When the mode is less than 6 the distribution would have a very short right tail. Figure 2.4 displays the tail probability $\Pr(Y > 16)$ for $Bes(\nu, a(\nu))$ and shows that $\Pr(Y > 16) \leq 2 \times 10^{-4}$ for $-1 < \nu \leq 1000$, a range of $\nu$ wide enough for our intended use in Bayesian analysis.

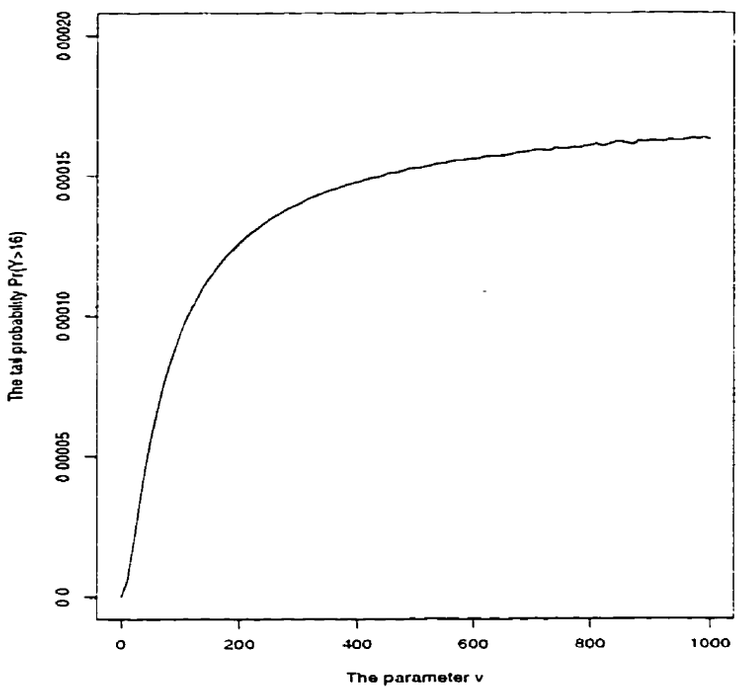


*Fig.* 2.4 The tail probability Pr $(Y > 16)$ for Bessel distribution $Bes(\nu, a(\nu))$.

We can easily argue that a Bessel distribution $Bes(\nu, a)$ with $a < a(\nu)$ has an even shorter right tail than that of $Bes(\nu, a(\nu))$. Suppose we have a sample $Y \sim Bes(\nu, a(\nu))$ and want to use this to generate a sample from $Bes(\nu, a)$ with $a < a(\nu)$. Applying von Neumann's rejection sampling theorem we see that $Y$ will be accepted if it is smaller than some random quantity. Therefore, $Bes(\nu, a)$ has a shorter right tail and $\Pr(Y > 16) \leq 2 \times 10^{-4}$ is valid for all $Bes(\nu, a)$ with $a \leq a(\nu)$, $-1 < \nu \leq 1000$. Consequently, table sampling is appropriate for the case $m(\nu, a) < 6$ which requires only a relative magnitude of the probability mass and evaluation of the Bessel function is not necessary. For $Bes(\nu, a)$, $m(\nu, a) < 6$, we set

$$q_0 = 1 \quad \text{and} \quad q_k = \frac{a^2}{4k(k+\nu)} q_{k-1}, \quad k = 1, \cdots, 16.$$

and then $s_k = \sum_{i=0}^{k} q_i / \sum_{i=0}^{16} q_i$. A sample $Y \sim Bes(\nu, a)$ can be drawn by generating a uniform random variable $\eta \sim U(0,1)$ and letting $Y = \sum_{k=0}^{16} I(\eta > s_k)$. Higher accuracy can be reached at the cost of sampling from a longer table. For example, the tail probability $\Pr(Y > 20) \leq 2.02 \times 10^{-6}$ is valid for all $Bes(\nu, a)$ with $a \leq a(\nu)$, $-1 < \nu \leq 1000$.

*CHAPTER 3*

# *Survival Analysis with Many Parameters*

## 3.1 The Infinite-Dimensional Gamma Prior

A prior distribution represents our knowledge about the unknown parameter before we look at the data. In the nonparametric survival problem, a natural way to summarize this knowledge is in terms of a shape parameter $\lambda_0(t) > 0$, which is our guess at the hazard function, and a confidence parameter $c > 0$ in the prior specification. The shape parameter summarizes our estimate of the unknown hazard function based on past experience. The confidence parameter specifies the degree of uncertainty we attribute to that estimate.

Ferguson (1973) remarked that a desirable prior for nonparametric problem should:

(i) have a large support on the space of probability distributions.

(ii) lead to a tractable posterior given the observations.

The Dirichlet process as a prior on the space of probability distributions meets these two requirements. But it assigns full probability to the class of discrete distributions, and this has caused difficulty and inconvenience. Therefore, a third desirable property might be

(iii) assign full probability to continuous distributions.

We begin by introducing a model with many parameters. Suppose that there is a small time unit $\delta > 0$ such that the class of hazard functions defined by

$$\lambda(t|\theta) = \theta_i \, \lambda_0(t), \quad t \in (i\delta - \delta, \, i\delta], \quad i = 1, 2, \cdots. \tag{3.1}$$

is rich enough for our purpose of describing the life distribution. Here, $\{\theta_i\}_{i=1}^{\infty}$ is a sequence of positive parameters and $\lambda_0(t)$ is the initial estimate for the hazard function. In a Bayesian framework we need to specify the joint probability distribution of $\{\theta_i\}_{i=1}^{\infty}$ in order to summarize the uncertainty in our knowledge about the unknown hazard function. As we pointed out in section 2.1, a Gamma prior is a natural choice in survival analysis. To make $\lambda_0(t)$ the prior mean, the marginal distribution of $\theta_i$ is chosen to be $G(c, c)$ where $c > 0$. But independence between the parameters would be unrealistic since $\delta$ is usually small, and the knowledge represented by $\theta_{i-1}$ and $\theta_i$ would often be strongly correlated.

We assume that $\{\theta_i\}_{i=1}^{\infty}$ constitutes a Markov chain. Under the constraint set by the marginal distributions, we can generate such a Markov chain by discretely sampling a modified squared Bessel process. Let $\xi(t)$ be a $2c$-dimensional squared Bessel process with $\xi(0) = 0$ and let $h$ be a strictly increasing function. We define the prior of $\theta_i$, $i = 1, 2, \cdots$, as the distribution of the Markov chain produced by sampling $\xi_h(t) = \xi(h(t))/(2h(t)c)$ at $i\delta$, $i = 1, 2, \cdots$. By doing so we have introduced an infinite-dimensional Gamma prior on hazard functions.

To specify such a prior we have used four items. The time unit $\delta$, the shape parameter $\lambda_0(t)$, the confidence $c$ and a function $h$ called the smoothing parameter. Each plays an important role in the prior. For simplicity we use the notation $IG(\delta, \lambda_0(t), c, h(t))$ to represent the prior constructed above. As stated before, $\delta$ determines the richness of the support of the prior and its choice depends on the precision required in the data analysis. In practice, data must be recorded within some precision limit, say, a year, a month, a day or even a minute, according to the nature of the investigation. This kind of data precision is determined at the experimental design stage; to avoid any loss of information, we can choose $\delta$ no larger than the precision limit. Marginally, the distribution of the hazard function $\lambda(t|\theta)$ is $G(c, c/\lambda_0(t))$. Our guess, $\lambda_0(t)$, is the mean and also the center of the distribution, and $c$, which determines the degree of concentration, is a measure of confidence. The larger $c$, the more the prior is concentrated. The smoothing parameter $h$ is designed to control the associations between $\theta_i$, $i = 1, 2, \cdots$. For $i < j$, the dependence coefficient between $\theta_i$ and

$\theta_j$ as defined in section 2.3 is $h(i\delta)/h(j\delta)$ so that $h$ determines the degree of associations between the $\theta_i$, $i = 1, 2, \cdots$. The choice of $h$ could be very arbitrary, but due to some technical difficulty we only discuss two special cases: the stationary case $h(t) \propto e^{\mu t}$ and the shape-dependent case $h(t) \propto \Lambda_0^{\mu}(t)$, where $\mu > 0$. In either case, note that, when $\mu \to \infty$, the Markov chain $\{\theta_i\}_{i=1}^{\infty}$ returns to an independent sequence. The smoothing parameter $h(t) \propto e^{\mu t}$ providing a stationary dependence structure on the sequence $\{\theta_i\}_{i=1}^{\infty}$ is the most uniform representation of the relationships between parameters. On the other hand, the shape-dependent smoothing parameter $h(t) \propto \Lambda_0^{\mu}(t)$, which measures the associations according to a kind of distance $[\Lambda_0(s)/\Lambda_0(t)]^{\mu}$, $s < t$ on the time, is also intuitively acceptable. More discussion of the smoothing parameter is provided in subsequent sections.

Two extreme cases in confidence should be discussed separately: $c \to \infty$ and $c \to 0$. Obviously, as $c \to \infty$, $IG(\delta, \lambda_0(t), c, h(t))$ becomes concentrated at its mean and no uncertainty exists in the prior knowledge. As $c \to 0$, the prior $IG(\delta, \lambda_0(t), c, h(t))$ approaches a non-informative prior under which the $\log \theta_i$ are independent and uniformly distributed over the real line. More discussion of this is presented later in this chapter.

As an ideal situation we may consider the exact failure times where $\delta \to 0$, and the stochastic process $\lambda_0(t)\xi_h(t)$ can be viewed as a kind of limit of the hazard function defined above. The parameter in this case should be $\theta_t = \xi_h(t)$, a continuous time process. However, mathematical analysis falls short in dealing with a general smoothing parameter. As it turns out, we must confine our study to the shape-dependent smoothing parameter. As an illustration, we work with $h(t) = \sqrt{\Lambda_0(t)}$ for the exact failure time data. The hazard function in this case can be represented as $\lambda(t|\theta) = \xi(\gamma(t))\gamma'(t)$ where $\gamma(t) = \sqrt{\Lambda_0(t)/c}$, and the cumulative hazard is thus an integral transform of the squared Bessel process. Some moments of $\Lambda(t|\theta)$ are available. For example, the prior mean and variance of the cumulative hazard can be obtained from the generalized Cameron-Martin formula (2.13):

$$E\,\Lambda(t|\theta) = \Lambda_0(t) \quad \text{and} \quad \text{var}\,\Lambda(t|\theta) = \frac{1}{6c}\Lambda_0^2(t).$$

From this we see that the shape and confidence control the features of the prior.

## 3.2 The Posterior Under Censored Data

We first solve the posterior problem for life table data. Let $t_i\delta, i = 1, \cdots, n$ be the observed failure times and $d_i$ be the number of subjects failing at $t_i$. Without loss of generality we assume that $t_i$ are all integers. Let $s_j$ denote the number of subjects at risk just before $j\delta$ and $N$ denote the largest index for which $s_N$ is non-zero. The prior density for $\theta_j, j = 1, \cdots, N$, is proportional to

$$\prod_{j=1}^{N} q(h(j\delta) - h(j\delta - \delta), 2h(j\delta - \delta)c\theta_{j-1}, 2h(j\delta)c\theta_j)$$

where $q$ is defined in (2.8) with $\nu = c - 1$ and $\theta_0 = 0$. The likelihood for parameters $\theta_j, j = 1, \cdots, N$ is

$$\theta_{t_1}^{d_1} \cdots \theta_{t_n}^{d_n} \exp[-\sum_{j=1}^{N} s_j\theta_j(\Lambda_0(j\delta) - \Lambda_0(j\delta - \delta))]$$

and by Bayes theorem, the posterior density is proportional to

$$(\theta_1\theta_N)^{(c-1)/2}\theta_{t_1}^{d_1} \cdots \theta_{t_n}^{d_n} \exp(-\sum_{j=1}^{N} a_j\theta_j) \prod_{j=2}^{N} I_{c-1}(\frac{\sqrt{\theta_{j-1}\theta_j}}{b_j}) \tag{3.2}$$

where, for $j = 1, \cdots, N-1$,

$$a_j = c[\frac{h(j\delta)}{h(j\delta) - h(j\delta - \delta)} + \frac{h(j\delta)}{h(j\delta + \delta) - h(j\delta)}] + s_j[\Lambda_0(j\delta) - \Lambda_0(j\delta - \delta)],$$

$$a_N = c[\frac{h(N\delta)}{h(N\delta) - h(N\delta - \delta)}] + s_N[\Lambda_0(N\delta) - \Lambda_0(N\delta - \delta)],$$

and $$b_j = \frac{h(j\delta) - h(j\delta - \delta)}{2c\sqrt{h(j\delta - \delta)h(j\delta)}} \quad j = 2, \cdots, N.$$

When $h(t) \propto e^{\mu t}$, the expressions above can be simplified to

$$a_j = c \coth \frac{\mu\delta}{2} + s_j[\Lambda_0(j\delta) - \Lambda_0(j\delta - \delta)], \quad j = 1, \cdots, N-1,$$

$$a_N = \frac{1}{2}c(1 + \coth \frac{\mu\delta}{2}) + s_N[\Lambda_0(N\delta) - \Lambda_0(N\delta - \delta)],$$

$$b_j = \frac{1}{c} \sinh \frac{\mu\delta}{2} \quad j = 2, \cdots, N.$$

Gibbs sampling from (3.2) may not always be feasible for $N$ could be very large. The high dimension of this density function may cause slow convergence of the Markov chain generated by the Gibbs sampler. If this is the case, a possible approach is to break the parameter into two parts, $(\theta_1, \theta_{t_1}, \cdots, \theta_{t_n}, \theta_N)$, which will be sampled by the Gibbs sampler, and $\theta_j$, $j \in \{k : 1 < k < N, k \neq t_i\}$, viewed as nuisance parameters at this stage, though they may be equally important in later Bayesian inference. Integrating out the nuisance parameters can be done by using (2.9) repeatedly. To illustrate this we state a simple result.

LEMMA 3.1. *Let*

$$L(x, y) = \int_0^\infty \cdots \int_0^\infty \exp(-a_1\theta_1 - \cdots - a_n\theta_n) I_\nu(\frac{\sqrt{x\theta_1}}{b_1}) I_\nu(\frac{\sqrt{\theta_1\theta_2}}{b_2}) \cdots I_\nu(\frac{\sqrt{\theta_n y}}{b_{n+1}}) d\theta_1 \cdots d\theta_n$$

*for $a_i, b_i > 0$. Then, $L(x, y) \propto \exp(l_1 x + l_2 y) I_\nu(\sqrt{xy}/l_{12})$ where the coefficients $l_1$, $l_2$ and $l_{12}$ depend on only $a = (a_1, \cdots, a_n)$ and $b = (b_1, \cdots, b_{n+1})$. Furthermore, if we define $\{A_i\}_{i=1}^n$ and $\{B_i\}_{i=1}^n$ by $A_1 = a_1$, $B_1 = b_1$ and*

$$A_k = a_k - \frac{1}{4A_{k-1}b_k^2}, \quad B_k = 2A_{k-1}B_{k-1}b_k, \quad 2 \leq k \leq n,$$

*then*

$$l_1(a, b) = \sum_{k=1}^n \frac{1}{4A_k B_k^2}, \quad l_2(a, b) = \frac{1}{4A_n b_{n+1}^2}, \quad l_{12}(a, b) = 2A_n B_n b_{n+1}.$$

The sequences $\{A_i\}_{i=1}^n$ and $\{B_i\}_{i=1}^n$ give a step-by-step prescription for the calculation of the multiple integral using (2.9). After integrating out $\theta_1$, the roles of $a_2$ and $b_2$ are replaced by $A_2$ and $B_2$ respectively. Next, integrating out $\theta_2$ we see that $a_3$ and $b_3$ are replaced by $A_3$ and $B_3$. This process continued establishes the lemma.

Applying this lemma to the density (3.2), we can eliminate all the nuisance parameters. The marginal density of $(\theta_1, \theta_{t_1}, \cdots, \theta_{t_n}, \theta_N)$ can be expressed as

$$p(\theta_1, \theta_{t_1}, \cdots, \theta_{t_n}, \theta_N) \propto (\theta_1\theta_N)^{(c-1)/2} \theta_{t_1}^{d_1} \cdots \theta_{t_n}^{d_n}$$

$$\exp(-a_1'\theta_1 - \sum_{i=1}^n a_{t_i}'\theta_{t_i} - a_N'\theta_N)\, I_{c-1}(\frac{\sqrt{\theta_1\theta_{t_1}}}{b_{t_1}'}) \cdots I_{c-1}(\frac{\sqrt{\theta_{t_{n-1}}\theta_{t_n}}}{b_{t_n}'}) I_{c-1}(\frac{\sqrt{\theta_{t_n}\theta_N}}{b_N'}) \qquad (3.3)$$

where the constants $\{a'_j\}$ and $\{b'_j\}$ are easily calculated from $\{a_j\}$ and $\{b_j\}$ by functions $l_1$, $l_2$ and $l_{12}$ defined in the lemma. For example, when $2 \leq i \leq n-1$,

$$a'_{t_i} = a_{t_i} - l_1(a_{t_i+1}, \cdots, a_{t_{i+1}-1}; b_{t_i+1}, \cdots, b_{t_{i+1}}) - l_2(a_{t_{i-1}+1}, \cdots, a_{t_i-1}; b_{t_{i-1}+1}, \cdots, b_{t_i}).$$

$$b'_{t_i} = l_{12}(a_{t_{i-1}+1}, \cdots, a_{t_i-1}; b_{t_{i-1}+1}, \cdots, b_{t_i}). \quad 2 \leq i \leq n.$$

The Gibbs sampler can now be applied to density (3.3) and the detailed sampling scheme is given later in this section.

Suppose now we have a large sample of $(\theta_{t_1} \cdots \theta_{t_n})$ at hand. How can we explore the posterior of the full parameter? Specifically, given $\theta_{t_{i-1}}$ and $\theta_{t_i}$, $(\theta_{t_0} = 0)$ how can we generate a sample of $(\theta_{t_{i-1}+1}, \cdots, \theta_{t_i-1})$? One can easily see that this is equivalent to sampling from a density defined by the integrand in $L(x, y)$. The marginal density of $\theta_n$ is proportional to $\exp(-A_n\theta_n) I_{c-1}(\sqrt{x\theta_n}/B_n) I_{c-1}(\sqrt{\theta_n y}/b_{n+1})$ and thus it has a randomized Gamma of the second type. After obtaining a sample of $\theta_n$ we consider the conditional density of $\theta_{n-1}$ given $\theta_n$ which is proportional to $\exp(-A_{n-1}\theta_{n-1}) I_{c-1}(\sqrt{x\theta_{n-1}}/B_{n-1}) I_{c-1}(\sqrt{\theta_{n-1}\theta_n}/b_n)$, again a randomized Gamma distribution of the second type. Note that the characterization of the randomized Gamma distribution of the second type in section 2.3 gives a sampling scheme. Therefore, we can generate $(\theta_1, \cdots, \theta_n)$ as a chain from $\theta_n$ to $\theta_1$.

As stated before, the prior $IG(0, \lambda_0(t), c, \sqrt{\Lambda_0(t)})$ can be used to analyze the exact failure time data where a discrete time unit is not needed. Consider continuous lifetime data where each observation is either a failure or a right-censored time. Suppose that $t_0 = 0$ and we observed failures at time $t_1 < \cdots < t_n$ and $d_i$ subjects failed at $t_i$ ; censorings that occur in $[t_i, t_{i+1})$ are adjusted to $t_i$. It is generally agreed that an estimate of the cumulative hazard function is more easily interpreted than an estimate of the hazard function itself. Let $\Phi_i$ denote the increment of the cumulative hazard over $(t_{i-1}, t_i)$, $1 \leq i \leq n$. In the nonparametric case, the parameters $\Phi = (\Phi_1 \cdots, \Phi_n)$ are often the quantities of primary interest, but these increments alone are not enough to describe the probabilistic mechanism from which the data were generated. We have to introduce another parameter $\phi = (\phi_1, \cdots \phi_n)$ where $\phi_i$ is

the hazard at time $t_i$ divided by $\gamma'(t_i)$. Then the likelihood for $(\phi, \Phi)$ given the data can be expressed as

$$L(\phi, \Phi) = \phi_1^{d_1} \cdots \phi_n^{d_n} \exp(-s_1\Phi_1 - \cdots - s_n\Phi_n)$$

where $s_i$ denotes the number of subjects at risk just before time $t_i$.

Under the prior $IG(0, \lambda_0(t), c, \sqrt{\Lambda_0(t)})$, the joint prior density $\pi(\phi, \Phi)$ can be written as the product of the density $\pi(\phi)$ and the conditional density $\pi(\Phi|\phi)$. Then, from the transition density of the squared Bessel process,

$$\pi(\phi) \propto \prod_{i=1}^{n} q(\gamma(t_i) - \gamma(t_{i-1}), \phi_{i-1}, \phi_i),$$

where $\phi_0 = 0$ and again, by the Markov property,

$$\pi(\Phi|\phi) = \prod_{i=1}^{n} \pi(\Phi_i \mid \phi_{i-1}, \phi_i).$$

Hence, the posterior density $p(\phi, \Phi)$ is proportional to

$$\phi_1^{d_1} \cdots \phi_n^{d_n} \pi(\phi) \prod_{i=1}^{n} \exp(-s_i\Phi_i)\pi(\Phi_i \mid \phi_{i-1}, \phi_i),$$

and the marginal density $p(\phi)$ can be obtained by integrating out $\Phi$. Note that,

$$\begin{aligned} & \int \exp(-s_i\Phi_i)\pi(\Phi_i \mid \phi_{i-1}, \phi_i)d\Phi_i \\ = \; & E\left\{(\exp[-s_i \int_{\gamma(t_{i-1})}^{\gamma(t_i)} \xi(t)dt] \mid \xi(\gamma(t_{i-1})) = \phi_{i-1}, \xi(\gamma(t_i)) = \phi_i)\right\}. \end{aligned}$$

From path integral theory we have

$$\int \exp(-s_i\Phi_i)\pi(\Phi_i \mid \phi_{i-1}, \phi_i)d\Phi_i = \psi_{s_i}(\gamma(t_i) - \gamma(t_{i-1}), \phi_{i-1}, \phi_i),$$

where $\psi$ is defined by (2.14).

It follows that the posterior density function of $\phi$ is

$$p(\phi) \propto (\phi_1\phi_n)^{(c-1)/2}\phi_1^{d_1} \cdots \phi_n^{d_n} \exp(-a_1\phi_1 \cdots - a_n\phi_n) \prod_{i=2}^{n} I_{c-1}(\frac{\sqrt{\phi_{i-1}\phi_i}}{b_i}). \qquad (3.4)$$

where $a_i$, $b_i$ are given by:

$$a_i = \frac{\sqrt{2s_i}}{2}c_i + \frac{\sqrt{2s_{i+1}}}{2}c_{i+1}. \quad \text{and} \quad b_i = \frac{\sinh\sqrt{2s_i}(\gamma(t_i) - \gamma(t_{i-1}))}{\sqrt{2s_i}} \quad i = 1, \cdots, n.$$
$$c_i = \coth\sqrt{2s_i}(\gamma(t_i) - \gamma(t_{i-1})) \quad i = 1, \cdots, n \ (c_{n+1} = 0).$$

We can see that the posterior distribution of $\phi$ is also a randomized Gamma distribution, but the mechanism of this randomization is very complex. Direct sampling from the posterior density such as (3.3) or (3.4) seems quite hard. Fortunately, the Gibbs sampler can help in this situation. Note that, the density (3.4) can be viewed as a marginal density of

$$\phi_1^{c+d_1-1} \cdots \phi_n^{c+d_n-1} \exp(-a_1\phi_1 \cdots - a_n\phi_n) \prod_{i=2}^{n} \frac{1}{\Gamma(r_i+1)\Gamma(r_i+c)} (\frac{\sqrt{\phi_{i-1}\phi_i}}{2b_i})^{2r_i}$$

by integrating or summing out the variables $r_i$. In this case there is an advantage in working with the joint density of $\phi$ and $r_i$ with $r_i$ as auxiliary variables. Gibbs sampling between $\phi = (\phi_1, \cdots, \phi_n)$ and $r = (r_1, \cdots, r_{n+1})$ where $r_1 = r_{n+1} = 0$ can be easily done as follows: Given $r$, we generate independent variables $\phi_i \sim G(c+d_i+r_i+r_{i+1}, a_i)$. Given $\phi$, we generate independent variables $r_i \sim Bes(c-1, \sqrt{\phi_{i-1}\phi_i}/b_i)$, $2 \le i \le n$. This is a very efficient way to sample the posterior with the form given by (3.4) and the convergence is rapid.

The conditional posterior of the hazard function given $\phi$ is described by a Bessel bridge. We have found that the conditional posterior distribution of $\lambda(t|\theta)$, $t_{i-1} \le t \le t_i$, $(t_0 = 0)$ given $\phi_{i-1}$ and $\phi_i$ is the same as that of the re-scaled Bessel bridge $\xi_{x_{i-1},x_i}[h_i(t)]\gamma'^2(t)/h_i'(t)$ where $x_{i-1} = \phi_{i-1}h_i'(t_{i-1})/\gamma'(t_{i-1})$, $x_i = \phi_i h_i'(t_i)/\gamma'(t_i)$ and

$$h_i(t) = \frac{\exp(2\sqrt{2s_i}\,\gamma(t)) - \exp(2\sqrt{2s_i}\,\gamma(t_{i-1}))}{\exp(2\sqrt{2s_i}\,\gamma(t_i)) - \exp(2\sqrt{2s_i}\,\gamma(t_{i-1}))} \quad i = 1, \cdots, n. \tag{3.5}$$

For $t \ge t_n$, $\lambda(t|\theta)$ remains a $2c$-dimensional squared Bessel process under time change $\gamma$ and scale change $\gamma'$ with initial state given by the posterior of $\phi_n$. These $n+1$ components are conditionally independent given $\phi$.

The derivation of this result is similar to the technique we just used. For $t_{i-1} = u_0 < u_1 < \cdots < u_k < t_i$, we add some components to the parameter $(\phi, \Phi)$. Let $\phi_j'$ and $\Phi_j'$ be

the hazard at $u_j$ and the increment of the cumulative hazard over $(u_{j-1}, u_j)$. Using the same technique, we can obtain the posterior density for $(\phi_1, \cdots, \phi_{i-1}, \phi'_1, \cdots, \phi'_k, \phi_i, \cdots, \phi_n)$ from which the conditional density of $(\phi'_1, \cdots, \phi'_k)$ given $\phi_{i-1}$ and $\phi_i$ is easily found. Finally, by straightforward algebra, we can compare this conditional density with that of a squared Bessel bridge. They have the same structure except for the difference in scales for which the transform (3.5) is introduced.

It is unfortunate that the marginal density for $\Phi$ is beyond our reach. Even the conditional density $p(\Phi_i|\phi_{i-1}, \phi_i)$ is generally too complicated to work with. But the corresponding Laplace transform can be obtained through this conditional posterior and path integral (2.19).

For any $\lambda > 0$,

$$\begin{aligned} & E\,(\exp(-\lambda\Phi_i)|\phi_{i-1}, \phi_i) \\ = \; & E\exp[-\lambda\int_{t_{i-1}}^{t_i} \frac{\xi_{x_{i-1},x_i}[h_i(t)]\phi'^2(t)}{h'_i(t)}\,dt] \\ = \; & E\exp[-\frac{\lambda}{8s_i}\int_0^1 \frac{\xi_{x_{i-1},x_i}(u)}{(u+p)^2}du] \\ = \; & E\{\exp[-\frac{\lambda}{8s_i}\int_0^1 \frac{\xi(u)}{(u+p)^2}du] \mid \xi(0) = 2\sqrt{2s_i}\,p\,\phi_{i-1},\ \xi(1) = 2\sqrt{2s_i}\,(1+p)\,\phi_i\}. \end{aligned}$$

where

$$p = \frac{\exp(2\sqrt{2s_i}\,\gamma(t_{i-1}))}{\exp(2\sqrt{2s_i}\,\gamma(t_i)) - \exp(2\sqrt{2s_i}\,\gamma(t_{i-1}))}.$$

Let $\tau_i = \sqrt{2s_i}[\gamma(t_i) - \gamma(t_{i-1})]$. The above Laplace transform can now be expressed as the product of

$$\exp[\frac{\sqrt{2s_i}}{2}(\phi_{i-1}+\phi_i)(\coth\tau_i - \sqrt{1+\lambda/s_i}\coth\sqrt{1+\lambda/s_i}\,\tau_i)] \tag{3.6}$$

and

$$\frac{\sqrt{1+\lambda/s_i}\sinh\tau_i}{\sinh\sqrt{1+\lambda/s_i}\,\tau_i}\, I_{c-1}(\frac{\sqrt{2(\lambda+s_i)\phi_{i-1}\phi_i}}{\sinh\sqrt{1+\lambda/s_i}\,\tau_i}) I_{c-1}^{-1}(\frac{\sqrt{2s_i\phi_{i-1}\phi_i}}{\sinh\tau_i}). \tag{3.7}$$

Thus, from section 2.4 we conclude that, given $\phi_{i-1}$ and $\phi_i$, a sample of $\Phi_i$ can be drawn as follows:

(i) *generate an independent random sequence* $r_{1,k}$, $k \geq 1$, *where* $r_{1,k}$ *is from a Poisson distribution with mean*

$$\frac{\sqrt{2s_i}}{\tau_i}(\phi_{i-1}+\phi_i)\frac{k^2\pi^2}{k^2\pi^2+\tau_i^2};$$

(ii) *generate a random variable* $r_2 \sim Bes(c-1, \sqrt{\phi_{i-1}\phi_i}/b_i)$ *independent of* $r_{1,k}$, $k \geq 1$;

(iii) *generate an independent random sequence* $\eta_k \sim G(c+r_{1,k}+2r_2, s_i/\tau_i^2)$, $k \geq 1$;

(iv) $\Phi_i = \sum_{k=1}^{\infty} \eta_k/(k^2\pi^2+\tau_i^2)$.

## 3.3 Bayesian Estimation

If our interest is only Bayesian estimation, the computation could be easier for only some numerical features of the posterior, such as the mode or moment, are needed.

As is well known, taking the posterior mode as a Bayesian estimate is numerically equivalent to the penalized likelihood approach. To estimate the parameters $\theta_j$, $j = 1, \cdots, N$ we consider the penalized likelihood (3.2) which can be viewed as a marginal likelihood from

$$\theta_1^{c+\beta_1-1}\cdots\theta_N^{c+\beta_N-1}\exp(-a_1\theta_1-\cdots-a_N\theta_N)\prod_{j=2}^{N}\frac{1}{\Gamma(r_j+1)\Gamma(r_j+c)}(\frac{\sqrt{\theta_{j-1}\theta_j}}{2b_j})^{2r_j}$$

by integrating or summing out the variables $r_j$, $2 \leq j \leq N$, where $\beta_j = d_i$ if $j\delta$ is failure time $t_i$ and zero otherwise.

For $c > 1$, the mode can be computed by the EM algorithm. If $r_j$, $2 \leq j \leq N$ are known the maximum penalized likelihood estimate for $\theta_j$ would have a closed form expression

$$\hat{\theta}_j = (c+\beta_j+r_j+r_{j+1}-1)/a_j$$

where $r_1 = r_{N+1} = 0$. On the other hand, given the current value of $\theta_j$, $1 \leq j \leq N$, $r_j$, $2 \leq j \leq N$ are conditionally independent with

$$r_j \sim Bes(c-1, \frac{\sqrt{\theta_{j-1}\theta_j}}{b_j}).$$

Therefore, the EM algorithm is appropriate for this situation which yields the following iterative procedure: if the current value of $\theta_j$ is $\theta_j^{(k)}$, the next value can be computed as

$$\theta_j^{(k+1)} = (c + \beta_j + \frac{\sqrt{\theta_{j-1}^{(k)}\theta_j^{(k)}}}{2b_j} R_{c-1}(\frac{\sqrt{\theta_{j-1}^{(k)}\theta_j^{(k)}}}{b_j}) + \frac{\sqrt{\theta_j^{(k)}\theta_{j+1}^{(k)}}}{2b_{j+1}} R_{c-1}(\frac{\sqrt{\theta_j^{(k)}\theta_{j+1}^{(k)}}}{b_{j+1}}) - 1)/\alpha_j.$$

The convergence of this algorithm is usually more rapid than that of a Gibbs sampler.

The role of the confidence and smoothing parameters can be further explained in the light of penalized likelihood. Given $r_j$, $2 \le j \le N$, the logarithm of the penalized likelihood (3.2) can be expressed as

$$\text{log-likelihood} + \sum_{j=1}^{N} (c + r_j + r_{j+1} - 1) \log \theta_j - \sum_{j=1}^{N} c[\frac{h(j\delta)}{h(j\delta) - h(j\delta - \delta)} +$$

$$+ \frac{h(j\delta)}{h(j\delta + \delta) - h(j\delta)}] \theta_j$$

where $h(N\delta + \delta)$ is taken as infinity. We see that $c$ and $h$ determine the way to modify the likelihood. The confidence $c$ is a global feature which controls the overall amount of modification; while $h$ describes the details of the modification for each of the components $\theta_j$. As a special case, the smoothing parameter $h(t) = e^{\mu t}$ treats each component equally.

Next, we show how to estimate the hazard function with the posterior (3.4) under quadratic loss. For $t_{i-1} \le t \le t_i$, the posterior conditional expectation of $\lambda(t|\theta)$ given $\phi_{i-1}$ and $\phi_i$ can be calculated by the property of squared Bessel bridge given in section 2.3. Then the posterior mean of the hazard function would be

$$\gamma'(t) Q_i(\frac{\sinh \sqrt{2s_i}(\gamma(t_i) - \gamma(t))}{\sinh \sqrt{2s_i}(\gamma(t_i) - \gamma(t_{i-1}))}, \frac{\sinh \sqrt{2s_i}(\gamma(t) - \gamma(t_{i-1}))}{\sinh \sqrt{2s_i}(\gamma(t_i) - \gamma(t_{i-1}))}).$$

where $Q_i(x, y) = A_{i-1}x^2 + 2B_i xy + A_i y^2$ with

$$A_i = E\,\phi_i, \qquad i = 0, \cdots, n. \tag{3.8}$$

$$B_i = b_i c + E\,\sqrt{\phi_{i-1}\phi_i} R_{c-1}(\frac{\sqrt{\phi_{i-1}\phi_i}}{b_i}), \qquad i = 1, \cdots, n. \tag{3.9}$$

while for $\lambda(t|\theta)$, $t > t_n$ the posterior mean is

$$\gamma'(t)[2c(\gamma(t) - \gamma(t_n)] + A_n].$$

where the expectation is taken with respect to the posterior of $\phi$.

Therefore, estimating the whole curve of the hazard function is equivalent to estimating $2n$ parameters or coefficients. If a sample of $\phi$ from (3.4) is available, this can be easily done. Subsequently, the Bayesian estimate of $\Phi$ is obtained as well.

## 3.4 Numerical Illustration

We illustrate our method by analyzing the survival of lung cancer patients using data reported by Prentice (1974). The data, which consist of survival times in days, are given in Table 3.1.

**Table 3.1**

VETERAN'S ADMINISTRATION LUNG CANCER TRIAL DATA

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group I | 8. | 10. | 11. | 25*. | 42. | 72. | 82. | 100*. | 110. | 118. |
| | 126. | 144. | 228. | 314. | 411 | | | | | |
| Group II | 4 | 7 | 10 | 13 | 16 | 18 | 18 | 20 | 21 | 22 |
| | 27 | 30 | 31 | 51 | 52 | 54 | 54 | 56 | 59 | 63 |
| | 97* | 117 | 122 | 123* | 139 | 151 | 153 | 287 | 384 | 392 |
| Group III | 3 | 8 | 12 | 35 | 92 | 95 | 117 | 132 | 162 | |
| Group IV | 12 | 100 | 103 | 105 | 143 | 156 | 162 | 177 | 182* | 200 |
| | 216 | 250 | 260 | 278 | 553 | | | | | |

* Censored

First we use the Markov chain prior with time unit one day or $\delta = 1$. Suppose that based on experience we have a shape specification $\Lambda_0(t) = 0.012\,t$ for the prior. We consider three levels of confidence ($c = 0.5,\ 10,\ 50$) and of smoothing parameter ($h_1(t) = \sqrt[4]{\Lambda_0(t)}$, $h_2(t) =$

$\sqrt{\Lambda_0(t)}$, $h_3(t) = \Lambda_0(t)$). For the moment, we use the smoothing parameter $h_2$ and we perform Bayesian analyses at different levels of confidence. The Gibbs sampler is used to obtain a large sample from the posterior of $\theta_{t_i}$, $i = 1, \cdots, n$.

Next, samples of $(\theta_{t_{i-1}+1}, \cdots, \theta_{t_i-1})$, $i = 1, \cdots, n$ are generated by the iterative procedures given in section 3.2 and, once this is available, Bayesian inference can easily be done. For example, under quadratic loss we use the posterior mean to estimate the hazard function. We can also transform the sample and compute the posterior mean for the survival function.

The subsequent discussion including all the figures in this section is regarding to the one-sample analysis for the Group I in the lung cancer data. In Figure 3.1 we compare the survival estimates for different values of confidence. It shows how the Bayesian analysis depends on the confidence. When $c$ is small, the estimate is close to the Kaplan-Meier estimate, the empirical result. As $c$ increases, the Bayesian estimate changes gradually from the empirical to the prior estimate $\exp(-0.012\,t)$.

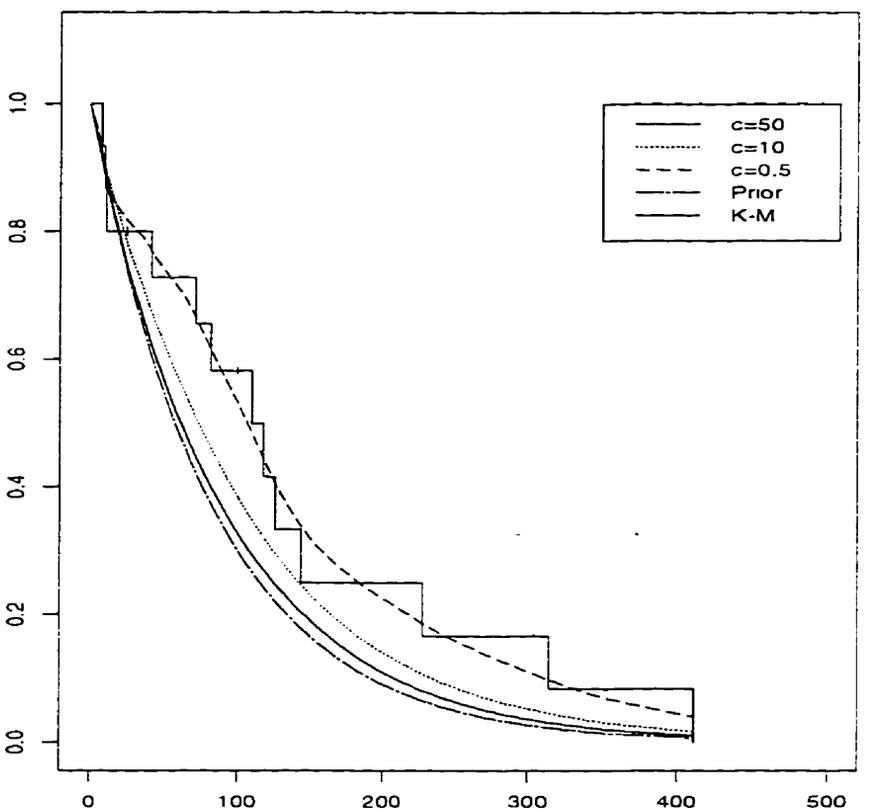


*Fig.* 3.1 Comparison of the Bayesian estimates under priors $IG(1, \Lambda_0, c, h_2)$, $c = 0.5, 10, 50$ with the prior estimate and Kaplan-Meier estimate for the survival curve.

The effect of the smoothing parameter can be examined by changing it from the very smooth $h_1(t) = \sqrt[4]{\Lambda_0(t)}$ to the moderate smooth $h_2(t) = \sqrt{\Lambda_0(t)}$ and then to the least smooth $h_3(t) = \Lambda_0(t)$ but keeping the confidence level $c = 0.5$. The estimated survival curves are all close to the Kaplan-Meier due to the low confidence.

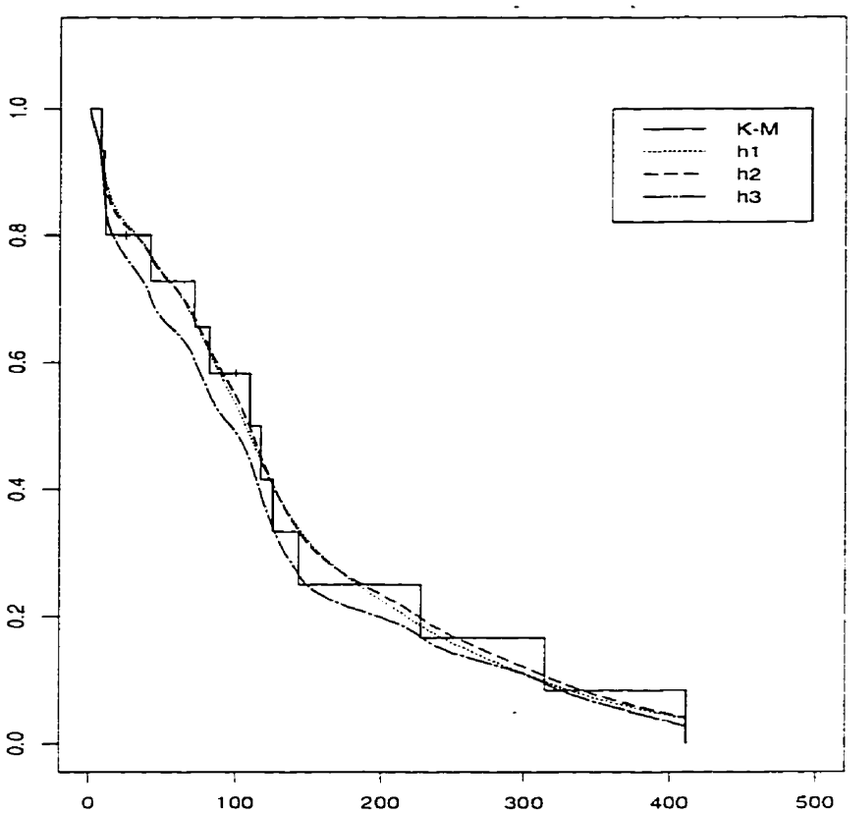


*Fig*. 3.2 Comparison of the Bayesian estimates of the survival function under priors $IG(1, \Lambda_0, 0.5, h_i)$, $i = 1, 2, 3$ with Kaplan-Meier estimate.

From this point of view the smoothing parameter does not affect the result much since the survival curve is quite stable when the smoothing parameter varies over a certain range. We also noticed that the curves corresponding to $h_1$ and $h_2$ are smoother than that corresponding to $h_3$, though the difference is not serious.

Figure 3.3 gives another view of the smoothing parameter. Since the confidence is low, the Bayesian estimate of the hazard is sensitive to the data. Thus, at a failure time the estimate is

large and yields a peak, with a valley between failures. Consequently, the Bayesian estimates are wave-like curves. But there is an obvious difference in the amplitudes of these waves, indicating that the degree of oscillation depends on the smoothing parameter in agreement with intuition. The effect of the smoothing parameter on the cumulative hazard or survival function is less apparent because the integral transform itself is a smoothing procedure. The difference in smoothness is then diminished.

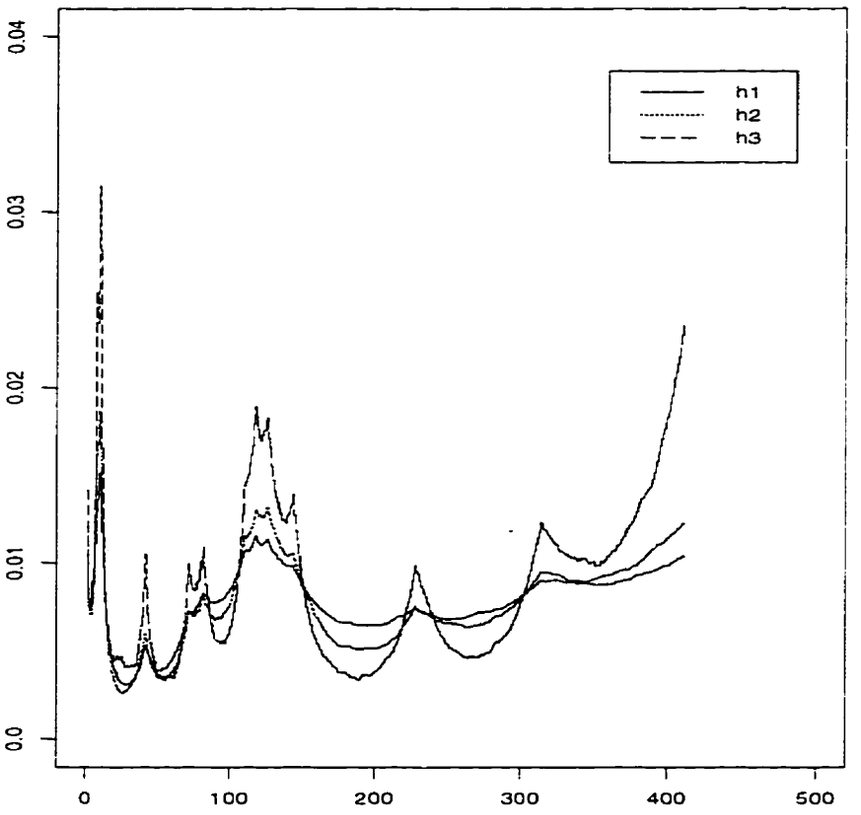


*Fig.* 3.3 Comparison of the Bayesian estimates of the hazard function under priors $IG(1, \Lambda_0, 0.5, h_i)$, $i = 1, 2, 3$.

The analysis of exact failure time data using the squared Bessel process prior is also

illustrated here. We first obtain the posterior density for $\phi$ and then a large sample from it. Next, we compute the coefficients from (3.9) and (3.10) and then the whole hazard curve. Three hazard curves are shown in Figure 3.4 corresponding to different confidence levels. These are smoothed piecewise since the means are calculated from different Bessel bridges over each interval between failures. Note that the prior mean is a horizontal line at the height 0.012. The effect of the confidence is to draw the curve towards that horizontal line and reduce the oscillation caused by the data.

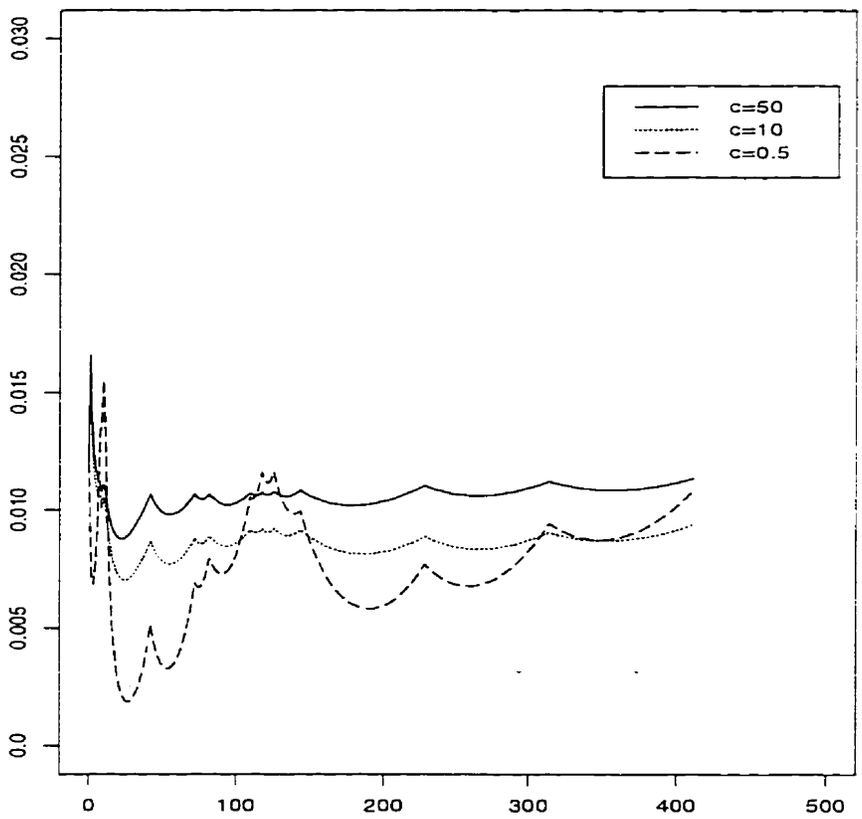


*Fig.* 3.4 Comparison of the Bayesian estimates of the hazard function under priors $IG(0, \Lambda_0, c, h_2)$, $c = 0.5,\ 10,\ 50$.

The integral transform of the hazard estimate would be the Bayesian estimate for the cumulative hazard. But inserting this Bayesian estimate of the cumulative hazard into the exponential function, we can only obtain an *ad hoc* Bayesian estimate for the survival function. Theoretically, we have to transform sample paths of the hazard into sample paths of the survival function and then to compute the mean. We have two choices here: either take the approximate approach to save computation, or connect the Bessel bridges to generate sample paths for the survival function. We display the *ad hoc* result in Figure 3.5 and note the similarity to the result reported in Figure 3.1 and discussed previously.

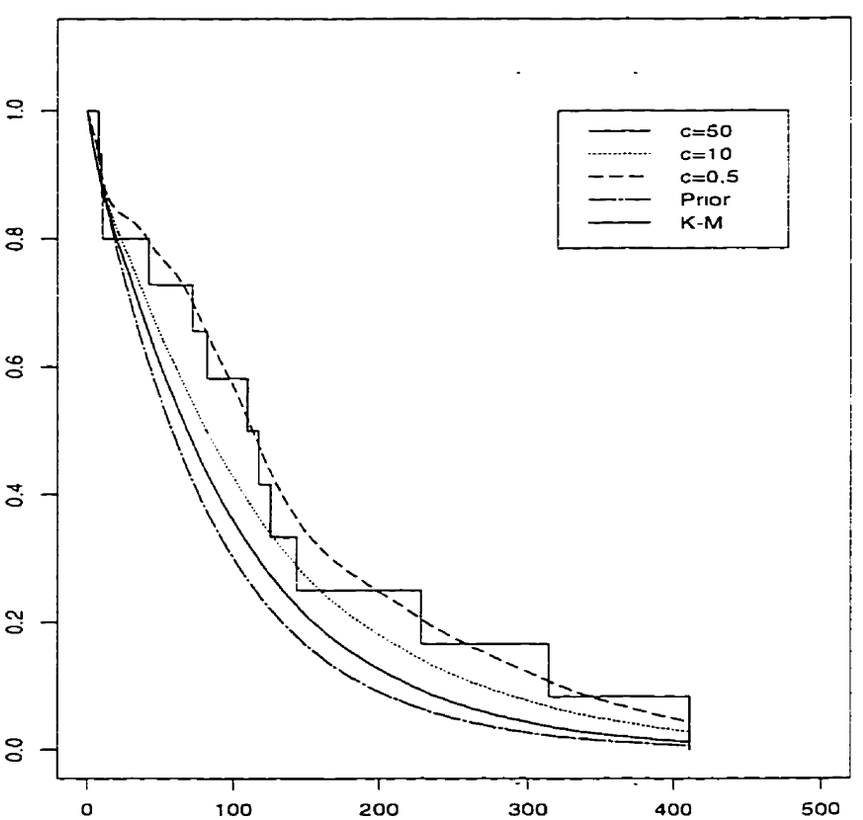


*Fig.* 3.5 Comparison of the Bayesian estimates under priors $IG(0, \Lambda_0, c, h_2)$, $c = 0.5, 10, 50$ with the prior estimate and Kaplan-Meier estimate for the survival function.

With the sample from the posterior we can certainly do more than estimation. For example, under the prior $IG(1, \Lambda_0, 10, h_2)$ we have 411 parameters and finally the posterior is numerically represented by a matrix of size 1000 rows and 411 columns in which each row is a sample of the parameters. It takes considerable computer memory storing such a posterior matrix. To visualize the posterior distribution, this matrix can be represented as sample paths of the survival function and be simultaneously plotted on a computer screen. For illustration, Figure 3.6 displays only a plot of 20 sample paths.

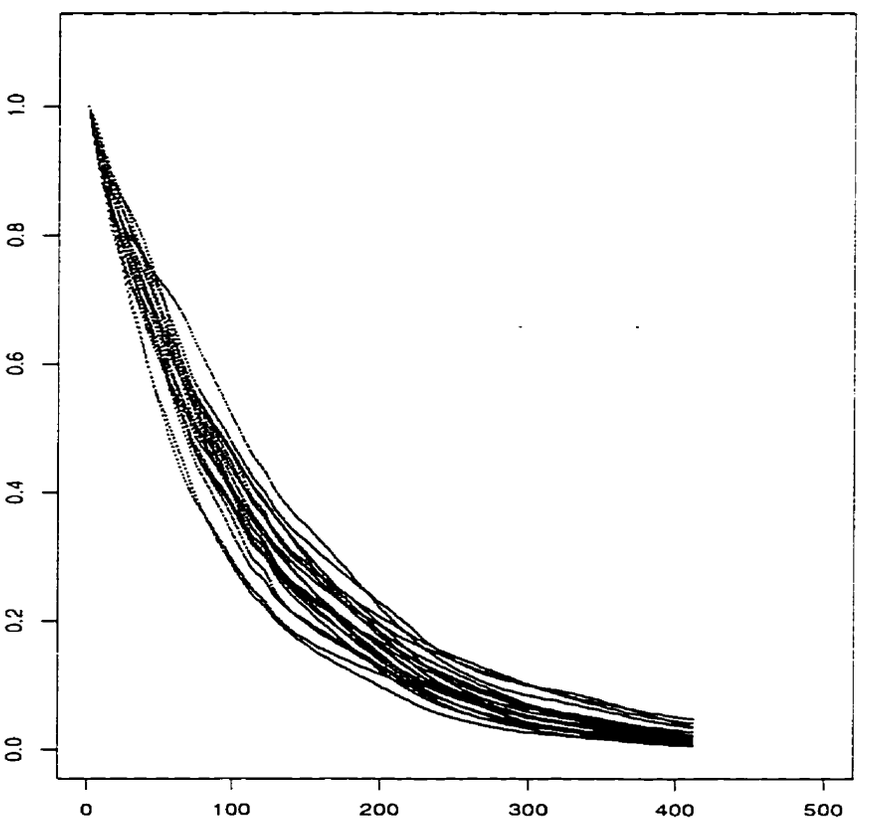


*Fig.* 3.6 Plot of twenty sample paths from the posterior of the survival function under the prior $IG(1, \Lambda_0, 10, h_2)$.

Our impression on the posterior of the survival function is gradually accumulated during the process of plotting. This might be the unique way to visually explore a posterior distribution with such a huge dimension. Unfortunately, one can only obtain an impression of the marginal distribution by looking at the final plot since it is impossible to distinguish one path from another.

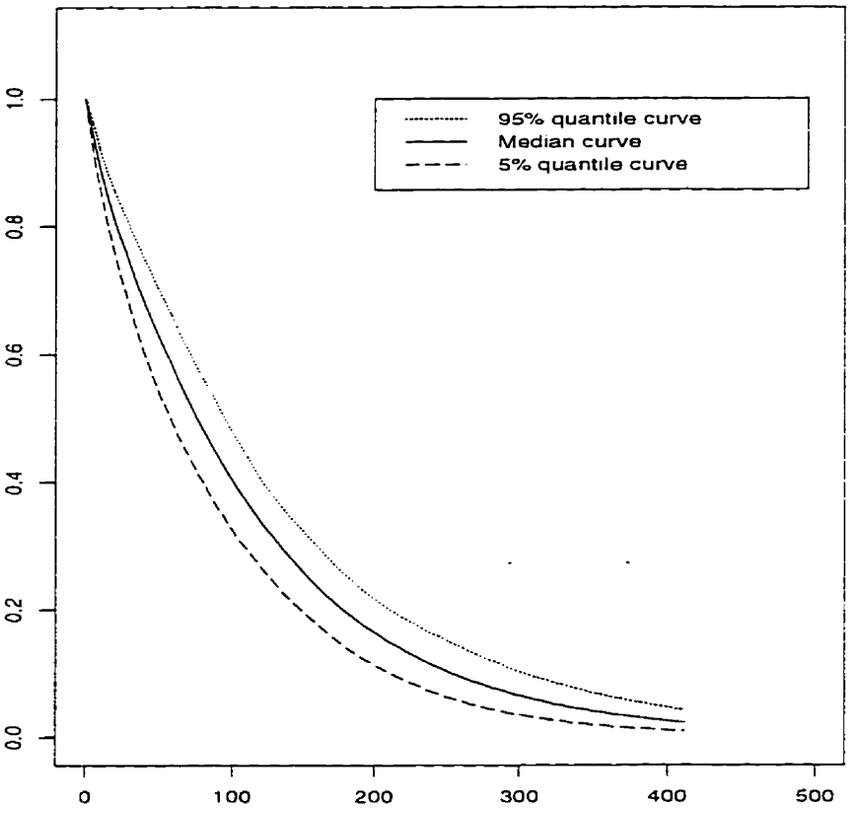


*Fig.* 3.7 Three quantile curves of the survival function under the prior $IG(1, \Lambda_0, 10, h_2)$.

One way to summarize the marginal distribution is to use each column of the matrix to obtain sample quantiles for the survival function at each grid point. Three quantile

curves are shown in Figure 3.7, where the median serves as the Bayesian estimate, and the other two form a 90% posterior probability interval for the survival function at a fixed time. Simultaneous coverage probability could also be determined numerically for these probability bands.

## 3.5 The Non-informative Prior

In a Bayesian analysis we always assume a prior distribution describing our initial knowledge about the parameter and then, use both the prior and data to derive the posterior. To some statisticians, the dependence of the results on the prior is completely unacceptable. Particularly in scientific investigations, it is hard to deny the need for standard posteriors which do not incorporate personal opinions. Suggestions have been made to solve this problem, and one possible technique is the so-called non-informative or reference prior which adds very little influence to the posterior. In other words, we would like to remove the subject information as much as possible but retain the Bayesian flavour in the inference.

There is no standard definition nor standard construction for a non-informative prior. Typical approaches are based on invariance (Jeffreys, 1967) or the limiting form of a conjugate prior (Novick, 1969), or the information-theoretic method of Lindley (1961) or Jaynes (1968). In many situations, the non-informative prior leads to a Bayesian analysis which agrees numerically with the classical results.

In terms of a non-informative prior for the model (3.1) we consider the limiting case where the confidence is extremely weak. The improper prior $IG(\delta, \lambda_0(t), 0, h(t))$ can be thought of as a limiting form of $IG(\delta, \lambda_0(t), c, h(t))$ as $c \to 0$. Recall from section 3.1, under $IG(\delta, \lambda_0(t), 0, h(t))$ the parameter sequence is mutually independent and each $\theta_j$ has a density proportional to $d\theta_j/\theta_j$. The posterior is immediately obtained that the parameter sequence is still independent and

$$\theta_{t_k} \sim G(d_i, s_{t_i}[\Lambda_0(t_i\delta) - \Lambda_0(t_i\delta - \delta)]) \quad \text{and} \quad \theta_j = 0 \quad \text{if} \quad j \neq t_k,\ 1 \leq k \leq n. \tag{3.10}$$

Hence, the hazard function is zero everywhere except in those intervals $(t_k\delta - \delta, t_k\delta]$, $k = 1, \cdots, n$. Based on this posterior, a Bayesian estimate of the cumulative hazard is

$$\frac{d_k}{s_{t_k}} r_k(t) + \sum_{i=1}^{k-1} \frac{d_i}{s_{t_i}}, \quad t_k\delta - \delta \le t < t_k\delta$$

where

$$r_k(t) = \frac{\Lambda_0(t) - \Lambda_0(t_k\delta - \delta)}{\Lambda_0(t_k\delta) - \Lambda_0(t_k\delta - \delta)}.$$

for $k = 1, \cdots, n$. The estimate is constant on each interval $[t_{k-1}\delta, t_k\delta - \delta)$.

Similarly, a Bayesian estimate for the survival function is

$$\hat{S}(t) = (1 - \frac{1}{s_{t_k} + r_k(t)})^{d_k} \prod_{i=1}^{k-1} (1 - \frac{1}{s_{t_i} + 1})^{d_i}, \quad t_k\delta - \delta \le t < t_k\delta, \tag{3.11}$$

and constant between $t_{k-1}\delta$ and $t_k\delta - \delta$, where an empty product is one. Note that the result still depends slightly on our initial knowledge. The data plays a main role in that the estimate at points $i\delta$ are independent of the prior: the shape parameter only provides a kind of interpolation.

Now, if we consider another limiting process in the posterior (3.11) by letting $\delta \to 0$, the hazard function becomes a linear combination of *Dirac* functions. The posterior of the cumulative hazard can be represented by

$$\sum_{t_i \le t} \frac{e_i}{s_{t_i}} \tag{3.12}$$

where $e_i \sim G(d_i, 1)$ are independent variables. The Bayesian estimate for the hazard based on (3.12) coincides exactly with the Nelson estimate (Nelson, 1972). Furthermore, the estimated survival curve would be

$$\hat{S}(t) = \prod_{t_i \le t} (1 - \frac{1}{s_{t_i} + 1})^{d_i}. \tag{3.13}$$

which is very similar but not identical to the Kaplan-Meier estimate. From (3.12) we can also develop a Bayesian interval for the survival curve by simulation. Applying this to the lung cancer data (Group I) we have the results displayed in Figure 3.8.

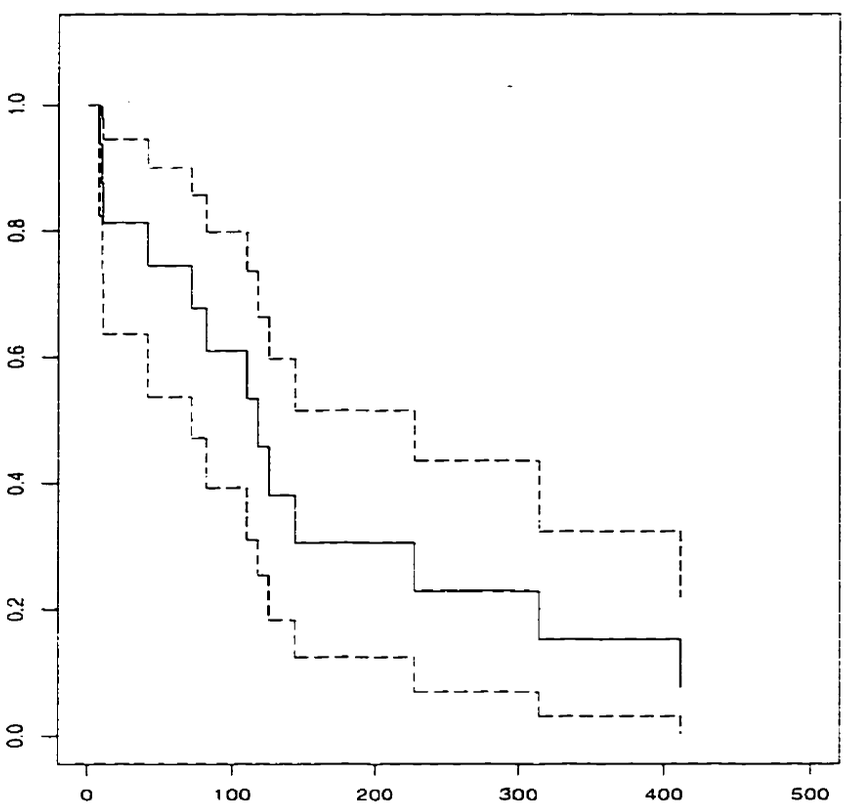


*Fig.* 3.8 The Bayesian point estimate compared with the 90% pointwise posterior probability intervals for the survival function for the lung cancer data (Group I) under a non-informative prior.

In fact, the estimate given by (3.13) has some advantage over the Kaplan-Meier, although they are equivalent asymptotically. The derivation of the Greenwood variance formula is not valid for small samples and the confidence interval is based on a normal approximation which may not be accurate. Actually, in some cases, the Greenwood lower bound curve could cross the zero line. On the other hand, the posterior (3.12) automatically provides the mean, the variance and the quantiles for each margin of the survival function.

We may also be interested in the non-informative case of the squared Bessel process prior. Let us first analyze the limiting behaviour of the Gibbs sampler for $\phi$. When $c \to 0$, the limiting distribution of $\phi_i$ will be independent. In fact, the Gibbs sampling scheme in section 3.3 has told the whole story. For convenience we follow the notations used there. As $c \to 0$ we have $a_i \to (\sqrt{2s_i} + \sqrt{2s_{i+1}})/2$ and $b_i \to \infty$ with an exponential rate. It is quite obvious that $r$ can be treated as zero and, under the limiting posterior, $\phi_i$, $i = 1, \cdots, n$ are independent and

$$\phi_i \sim G(d_i, \frac{\sqrt{2s_i} + \sqrt{2s_{i+1}}}{2}). \tag{3.14}$$

where $s_{n+1} = 0$.

Next, we calculate the limiting conditional distribution of $\Phi_i$ given $\phi_{i-1}$ and $\phi_i$ as $c \to 0$. According to section 3.2 the Laplace transform of this conditional distribution can be expressed as a product of two factors. It is easy to see that the first factor in the Laplace transform given by (3.6) tends to

$$\exp[-\frac{\sqrt{2s_i}}{2}(\phi_{i-1} + \phi_i)(\sqrt{1 + \lambda/s_i} - 1)].$$

But the limit of the second factor (3.7) cannot be directly evaluated since the arguments in the Bessel function also depend on $c$, and $I_{-1}(\cdot)$ is not well-defined by the power series given before. Let

$$x = \frac{\sqrt{2\phi_{i-1}\phi_i(1 + \lambda/s_i)}}{\sinh \sqrt{1 + \lambda/s_i}\,\gamma_i} \quad \text{and} \quad y = \frac{\sqrt{2s_i\phi_{i-1}\phi_i}}{\sinh \gamma_i}.$$

Then, by (2.2) the second factor can be written as

$$\frac{xI_{c-1}(x)}{yI_{c-1}(y)} = \frac{cI_c(x) + xI_{c+1}(x)}{cI_c(y) + yI_{c+1}(y)}$$

It is now obvious that the limit is one by eliminating the higher order infinitesimal terms. Therefore, we have

$$\lim_{c \to 0} E\,(\exp(-\lambda\Phi_i) \mid \phi_{i-1}, \phi_i) = \exp[-\frac{\sqrt{2s_i}}{2}(\phi_{i-1} + \phi_i)(\sqrt{1 + \lambda/s_i} - 1)]. \tag{3.15}$$

This posterior is in fact very informative that Bayesian estimates based on (3.14) and (3.15) are essentially different from conventional empirical results. For example, the cumulative hazard estimate

$$\sum_{t_i \le t} \frac{1}{2s_i} \left[ \frac{d_{i-1}\sqrt{s_i}}{\sqrt{s_{i-1}} + \sqrt{s_i}} + \frac{d_i \sqrt{s_i}}{\sqrt{s_i} + \sqrt{s_{i+1}}} \right], \quad (s_0 = s_{n+1} = \infty)$$

is almost half of the Nelson estimate.

Mathematically, this is not surprising. There are two limit operations: the confidence $c \to 0$ and the time unit $\delta \to 0$, whose order cannot be exchanged. It seems that taking the limit $c \to 0$ first and then $\delta \to 0$ leads to a result which agrees essentially with classical approaches. But, changing the order of these two limits might be dangerous.

From a Bayesian point of view, many problems arise in modeling ignorance. Even in the traditional parametric case, peculiar marginal behaviour of the posterior could occur (Dawid, Stone and Zidek, 1973) with an improper prior. It is also well known that a prior representing lack of knowledge about a parameter might, at the same time, incorporate strong knowledge about a transform of that parameter. Given a fixed time unit $\delta > 0$, both the hazard and cumulative hazard are reasonably represented by the prior. When our knowledge about the hazard is extremely vague, so is the knowledge about the cumulative hazard. However, as $c \to 0$, the Bessel process prior $IG(0, \Lambda_0, c, \sqrt{\Lambda_0})$ is non-informative about the hazard but very informative about the cumulative hazard. Consequently, when $c$ is extremely small, the posterior has been drawn in an unwanted direction.

The prior $IG(0, \Lambda_0, c, \sqrt{\Lambda_0})$ is viewed as a limiting case of $IG(\delta, \Lambda_0, c, \sqrt{\Lambda_0})$ when $\delta \to 0$. The numerical result shows this is true, at least for a confidence not too small. But when $c$ is extremely small, it is hard to tell whether the corresponding Bayesian analysis approximates that under $IG(\delta, \Lambda_0, c, \sqrt{\Lambda_0})$ with very small $\delta$, or approximates the poor result we just obtained. Therefore, the prior $IG(0, \Lambda_0, c, \sqrt{\Lambda_0})$ with extremely small $c$ should be avoided.

## 3.6 Choosing A Prior

To pure Bayesians the prior specification is virtually a subjective matter. But from the perspective of increasing data efficiency, information contained in related or historical datasets can sometime be incorporated into current analysis in the form of a prior distribution. For example, marketing and financial analysts use historical data to generate priors. In survival analysis, however, historical data may not always be available. In this section, we illustrate a way of choosing a prior from related datasets.

We choose the lung cancer data as an example. If we believe that similarity in survival exists within each treatment group, we may assume that the priors for the four subgroups in the standard treatment are independent and identically distributed and thus have common shape and confidence parameters. This is a typical empirical Bayes set-up.

To avoid heavy duty computation, we use some *ad hoc* rather than conventional estimation procedures. Intuitively, the Nelson estimate of the cumulative hazard function by pooling the data together is a good summary of the common aspects of survival in the four subgroups. The shape parameter can simply be taken as a smoothed function that fits the Nelson estimate well and has a positive first derivative.

In this particular example it is well approximated by a straight line $\Lambda_0(t) = 0.008t$ which can be taken as the shape parameter. The confidence should reflect the variation in the hazard plots and put most probability mass around the region formed by those curves. In practice several levels of confidence should be chosen and a sensitivity analysis is recommended.

There is also an *ad hoc* method of determining the confidence from the data. Suppose $\delta$ is the time unit for the whole dataset and the shape parameter $\Lambda_0$ is known. To assess the variation in the hazard plots we treat the hazard curves as realizations of the random process $\sum_j \theta_j[\Lambda_0(j\delta) - \Lambda_0(j\delta - \delta)]$. However, a difficulty arises that the random process must have a positive increment over a time interval $(j\delta - \delta, j\delta)$ but the plotted hazard curves are

pure jump functions. To proceed we have to spread the increments in the plotted hazard curves to smaller time intervals. For example, the increment of hazard over $(t_{i-1}\delta, t_i\delta)$ can be equally divided into increments over time intervals $(j\delta - \delta, j\delta)$, $j = t_{i-1} + 1, \cdots, t_i$, which is basically a smoothing procedure.

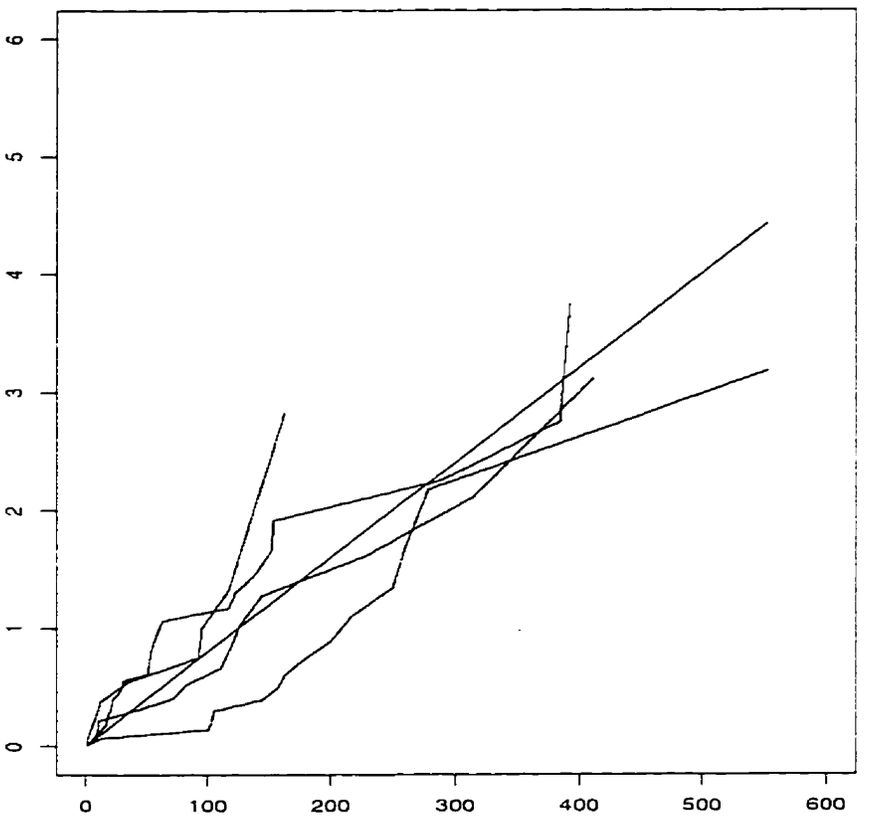


*Fig.* 3.9 Smoothed cumulative hazard plots of the four subgroups receiving standard treatment compared with the shape parameter $\Lambda_0(t) = 0.008\, t$.

If we think that the four increments over a given interval $(j\delta - \delta, j\delta)$ in the four hazard curves are independent realizations of $\theta_j[\Lambda_0(j\delta) - \Lambda_0(j\delta - \delta)]$, then the estimate $\hat{\theta}_j$ for $\theta_j$ is

easily obtained as an average. Next, since $\{\theta_j\}_{j=1}^{\infty}$ is assumed to be a Markov chain, under mild condition on the smoothing parameter $h$, the law of large numbers that

$$\frac{1}{m}\sum_{j=1}^{m}\theta_j^2 \;\to\; 1+\frac{1}{c} \qquad m\to\infty,$$

offers an estimate for the confidence. In the lung cancer data example this method gives $\hat{c}\simeq 0.9$.

As we pointed out in section 3.3, the effect of smoothing parameter is secondary to the choice of confidence. The stationary smoothing parameter is, in a certain sense, the most unbiased choice. But the use of the shape-dependent smoothing parameter is also intuitively acceptable. It is unfortunate that a finite dataset cannot provide the desired information on the smoothness of the hazard function. On the other hand, smoothness is needed for a reasonable incorporation of initial knowledge and a continuous modeling of the survival.

*CHAPTER 4*

# *Data-Dependent Prior and Fiducial Inference*

## 4.1 General Remarks

A rigorous Bayesian framework requires a fully specified prior distribution before the data become available. But, in some cases, the model is actually built after examining the data. For example, the Kaplan-Meier estimate is shown be to a maximum likelihood estimate where the parameters are set up with reference to the failure times. So, in this situation, how can we assign a prior distribution to the model parameters without even having a model? If we have examined the data, then how can we obtain a legitimate prior in the logical content.

It is also a fact that the operationally convenient prior may differ according to the parameter of interest. In survival analysis, the parameters of interest as given in section 3.2 can be determined only after the failure times have been observed. We have struggled to avoid this kind of problem by setting parameters almost everywhere because we do not know where the failures will be. Then we tried to work out the posterior for a special margin by integrating many "nuisance parameters", but we still need the "nuisance parameters" to complete the posterior computation and Bayesian inference. The practical merit of this approach is questionable given the cost due to the intensive computations. It seems that we have to choose between two evils, the computational burden if we specify the prior before the data as we have done in the last chapter; or the logical contradiction if we specify the prior after the data. The theme of this chapter is simplification and we take the latter approach.

In fact, any theory has to be seriously compromised in practice. We might have a perfect representation of our initial knowledge. But if it is not operationally convenient, we probably need to seek an alternative which approximates the original representation fairly well and also leads to simpler computations. Note that a random process can be approximated by a high-dimensional probability distribution. The difference between a random process and a high-dimensional probability is that, the former is a well-organized class of distributions satisfying Kolmogorov's consistency, a feature that may be completely irrelevant in terms of incorporating information.

We illustrate this by considering the simple situation that $t_1 < \cdots < t_n$ is an ordered sample of size $n$ from an unknown distribution $F$, upon which inference will be drawn. According to Ferguson (1973), a Dirichlet process $F(t)$ would be constructed expressing our initial knowledge about the data distribution. On the other hand, a compromised approach might assume that $F$ is constant between observations and has jump of size $\theta_i$ at $t_i$. The parameter space is now shrunk to $(n-1)$-dimensional space, namely that $\theta = (\theta_1, \cdots, \theta_n)$ satisfying $\theta_1 + \cdots + \theta_n = 1$. $\theta_i > 0$ are taken as parameters and $F(t|\theta) = \sum_{i=1}^{n} \theta_i I(t \geq t_i)$. This kind of parameterization is data-dependent for the location of the jumps is specified by the data. A Bayesian framework is now built by assigning a convenient prior to $\theta$, possibly a Dirichlet distribution which approximates the Dirichlet process prior in some sense. This is conceptually more straightforward than the Dirichlet process prior and leads to similar results.

A topic related to this is Fisher's fiducial approach. Fiducial inference is a kind of pivotal inference that leads to a probability description of uncertainty about the parameter. A pivotal is a function of the data and the parameter whose distribution is completely known. The fiducial argument transforms the probability statement about the pivotal quantity into probability statement about the parameter after the sample is drawn. Let us recall Fisher's original example on the sample variance of a normal distribution. If a sample of $n$ observations $x_1, \cdots, x_n$ has been drawn from a normal population with variance $\sigma^2$ and

$s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2/(n-1)$, then, it is well known that the quantity $s^2/\sigma^2$ has a $\chi^2$ distribution with $n-1$ degree of freedom. For any constants $a$, $b$ $(b > a > 0)$, we can always find the probability

$$\Pr\left( a < \frac{s^2}{\sigma^2} < b \right) = p$$

from the $\chi^2$ table but this probability statement is based on the so-called sampling distribution where $\sigma^2$ is fixed and $s^2$ is random. However, once the sample is available $s^2$ becomes fixed and the interpretation for the probability has to be changed. One may argue that this probability statement is only valid before the sampling and should be discarded after. Imagine that we are tossing a coin. We can make a probability statement about the outcome only before a tossing.

Fisher (1935) argued that the probability statement above induces a distribution on the parameter $\sigma^2$ given $s^2$ and thus the probability becomes a description of parameter uncertainty. Although this approach has been criticized since its introduction, it offers, at least in some cases, very reasonable results. According to Fisher (1935), fiducial probability differs from Bayesian inverse probability in logical content. But, numerically, the fiducial inference coincides with Bayesian under a certain kind of non-informative prior. For example, Pitman's estimate for location parameter can be derive from both the fiducial and Bayesian approaches.

We shall discuss the fiducial approach in a nonparametric setting where the parameter is the unknown distribution itself.

## 4.2 A Data-Dependent Prior and Its Posterior

We now introduce a data-dependent prior on the hazard function. We start with a guess $\Lambda_0(t)$ at the cumulative hazard, or equivalently $\lambda_0(t) > 0$ at the hazard function with a certain degree of confidence. Suppose that failures occur at times $t_1 < \cdots < t_n$ and $d_i$ subjects fail at $t_i$; censorings that occur in $[t_i, t_{i+1})$, $i = 0, \cdots, n$ are adjusted to $t_i$ where

$t_0 = 0$ and $t_{n+1} = \infty$. We assume that uncertainty exists only in a finite-dimensional margin of the hazard function at $t_i$, $i = 1, \cdots, n$, and, once this margin is given, the whole hazard function can be obtained by a kind of interpolation. In practice, we suppose that the hazard function over each interval $(t_{i-1}, t_i]$ can be parameterized as

$$\lambda(t|\theta) = \theta_i \lambda_0(t), \quad t \in (t_{i-1}, t_i] \tag{4.1}$$

where $\theta = (\theta_1, \cdots, \theta_n)$. The hazard function defined by (4.1) has fixed jumps at $t_i$, $i = 1, \cdots, n$.

Given the data described above, the likelihood function for $\theta$ is proportional to

$$\theta_1^{d_1} \cdots \theta_n^{d_n} \exp[-\sum_{i=1}^{n} s_i \theta_i (\Lambda_0(t_i) - \Lambda_0(t_{i-1}))]$$

where $s_i$ is the number of individuals at risk just before time $t_i$. According to this likelihood, the most practical prior for $\theta$ would be the conjugate prior with $\theta_i \sim G(c, c)$ being independent. This conjugate prior allows us to perform easy Bayesian analysis and to derive sensible Bayesian and non-Bayesian results such as the $A(n)$, the Nelson hazard estimate and even Cox's partial likelihood. We have many reasons to use this prior if a Bayesian analysis is needed.

It is also possible to use a non-conjugate prior such as the multivariate Gamma. Let $\xi(t)$ be a $2c$-dimensional squared Bessel process with $\xi(0) = 0$ and let $h$ be an increasing function. If we assume that the components of the parameter $\theta$ are generated by sampling a random process $\xi_h(t) = \xi(h(t))/(2h(t)c)$ at times $t_i$, then the posterior density of $\theta$ is

$$p(\theta) \propto (\theta_1 \theta_n)^{(c-1)/2} \theta_1^{d_1} \cdots \theta_n^{d_n} \exp(-a_1\theta_1 \cdots - a_n\theta_n) \prod_{i=2}^{n} I_{c-1}(\frac{\sqrt{\theta_{i-1}\theta_i}}{b_i}) \tag{4.2}$$

where the constants $\{a_i\}_{i=1}^n$, $\{b_i\}_{i=2}^n$, which carry information from the data as well as the prior, are given by:

$$a_i = c[\frac{h(t_i)}{h(t_i) - h(t_{i-1})} + \frac{h(t_i)}{h(t_{i+1}) - h(t_i)}] + s_i[\Lambda_0(t_i) - \Lambda_0(t_{i-1})], \quad i = 1, \cdots, n-1,$$

$$a_n = \frac{h(t_n)c}{h(t_n) - h(t_{n-1})} + s_n[\Lambda_0(t_n) - \Lambda_0(t_{n-1})]:$$
$$b_i = \frac{h(t_i) - h(t_{i-1})}{2c\sqrt{h(t_{i-1})h(t_i)}} \quad i = 2, \cdots, n.$$

For this density, the Gibbs sampling schemes proposed in section 3.2 can be applied without difficulty.

Under the conjugate prior, we have

$$a_i = c + s_i[\Lambda_0(t_i) - \Lambda_0(t_{i-1})], \quad i = 1, \cdots, n.$$
$$b_i = \infty \quad i = 2, \cdots, n.$$

and thus $\theta_i \sim G(c + d_i, a_i)$ are independent.

The limit of the proper prior as $c \to 0$ gives the non-informative prior having the form $\pi(\theta_1, \cdots, \theta_n) \propto d\theta_1 \cdots d\theta_n/\theta_1 \cdots \theta_n$. This coincides with the Jefferys' reference prior in this case. The corresponding posterior distribution is given by:

$$\theta_i \sim G(d_i, s_i[\Lambda_0(t_i) - \Lambda_0(t_{i-1})]), \ i = 1, \cdots, n$$

independently. The increment of the cumulative hazard over $(t_{i-1}, t_i)$ is thus a Gamma variable from $G(d_i, s_i)$ which is the same as the posterior (3.12). But an essential difference exists: according to the model (4.1), the increment is continuously distributed over the whole interval $(t_{i-1}, t_i)$; while the posterior (3.12) describes the increment as an impulse at time $t_i$.

## 4.3 Proportional Hazards Regression

In practice, it is important to incorporate some covariates in our survival model because the population under investigation can rarely be treated as homogeneous. It is also a basic goal of survival analysis to study the dependence of life time on explanatory variables. In

this section we extend the survival model (4.1) to the regression case. Suppose that a subject with covariates $z = (z_1, \cdots, z_k)'$ has hazard function

$$\lambda(t|z) = \exp(\beta' z)\lambda^*(t). \qquad t \geq 0. \tag{4.3}$$

where $\beta = (\beta_1, \cdots, \beta_k)'$ are the regression parameters and $\lambda^*(t)$ is the baseline hazard.

As before, let the observed failures be $t_1 < \cdots < t_n$ and suppose that censorings in $[t_i, t_{i+1})$ are all adjusted to $t_i$. Our primary concern in this case is the estimation of the regression parameter. If $\Lambda^*(t)$ is observable at the data points $t_i$ the likelihood for $\beta$ would be

$$L(\beta) = \prod_{i=1}^{n} e^{\beta' z_i} \exp[-s_i(\beta)(\Lambda^*(t_i) - \Lambda^*(t_{i-1}))]. \tag{4.4}$$

where $s_i(\beta) = \sum_{j \in \mathcal{R}(t_i)} \exp(\beta' z_j)$ and $\mathcal{R}(t_i)$ is the risk set just before time $t_i$. The point estimation problem corresponds to maximizing a parametric likelihood.

Cox (1972, 1975) proposed an approach which gives a kind of likelihood for $\beta$ in the absence of any knowledge about the baseline hazard. A later work of Kalbfleisch (1978) assumed a vague knowledge concerning the baseline hazard, which is represented by a Gamma process prior. Then a Bayesian way of eliminating nuisance parameters is adopted and a likelihood for $\beta$ is obtained by integrating out the baseline hazard.

If the conjugate data-dependent prior specified in the last section is assigned to the baseline hazard choosing $\Lambda_0$ and $c$ as the shape and confidence parameters, then the marginal likelihood would be

$$\prod_{i=1}^{n} \frac{e^{\beta' z_i}}{[s_i(\beta) + \dfrac{c}{\Lambda_0(t_i) - \Lambda_0(t_{i-1})}]^{c+1}}.$$

Two extreme cases are of interest: the parametric likelihood when $c \to \infty$ and Cox's partial likelihood when $c \to 0$. Thus the integrated likelihood gives all the intermediate analyses when $c$ goes from zero to infinity.

The pure Bayesian approach would also put a prior on the regression parameters. For example, one may assume a uniform prior for $\beta$ independent of the prior for the baseline

hazard. Generally, the posterior of $\beta$ is hard to sample and only when the covariates are properly coded is the Gibbs sampler applicable. For simplicity we only consider a Bayesian many-sample problem. Suppose the hazard functions of these samples are expressed as $\eta_k \lambda^*(t)$, $k = 1, \cdots, m$ with $\lambda^*(t)$ being a baseline hazard. Let $s_{ik}$ be the number of subjects at risk in the $k$-th sample just before time $t_i$ and $d_{ik}$ be the number of deaths at $t_i$ in the $k$-th sample. Under the prior, the regression parameters $\eta = (\eta_1, \cdots, \eta_m)'$ are independent of the baseline hazard. We assign a data-dependent prior to the baseline hazard as described in the last section, and assume $\eta_k$, $k = 1, \cdots, m$ are independent with $\log \eta_k$ uniformly distributed over the real line. The likelihood for $(\eta, \theta)$ is

$$\theta_1^{d_1} \cdots \theta_n^{d_n} \eta_1^{d_{\cdot 1}} \cdots \eta_m^{d_{\cdot m}} \exp[-\sum_{k=1}^{m} \sum_{i=1}^{n} \eta_k s_{ik} \theta_i (\Lambda_0(t_i) - \Lambda_0(t_{i-1}))].$$

where $d_{i\cdot} = \sum_{k=1}^{m} d_{ik}$ denotes the total failures at time $t_i$, and $d_{\cdot k} = \sum_{i=1}^{n} d_{ik}$ the total failures in the $k$-th sample.

A Gibbs sampling from the posterior of $(\eta, \theta)$ can be easily carried out when the non-informative prior for $\theta$ is used. Given $\eta$, we draw independent variables

$$\theta_i \sim G(d_{i\cdot}, \sum_{k=1}^{m} s_{ik} \eta_k [\Lambda_0(t_i) - \Lambda_0(t_{i-1})]), \quad i = 1, \cdots, n,$$

and given $\theta$, we draw independent variables

$$\eta_k \sim G(d_{\cdot k}, \sum_{i=1}^{n} s_{ik} \theta_i [\Lambda_0(t_i) - \Lambda_0(t_{i-1})]), \quad k = 1, \cdots, m.$$

We illustrate this using the lung cancer data presented in section 3.4 comparing the four subgroups receiving standard treatment. We can arbitrarily choose a shape parameter, say $\Lambda_0(t) = 0.008t$, which has actually no influence on the comparison.

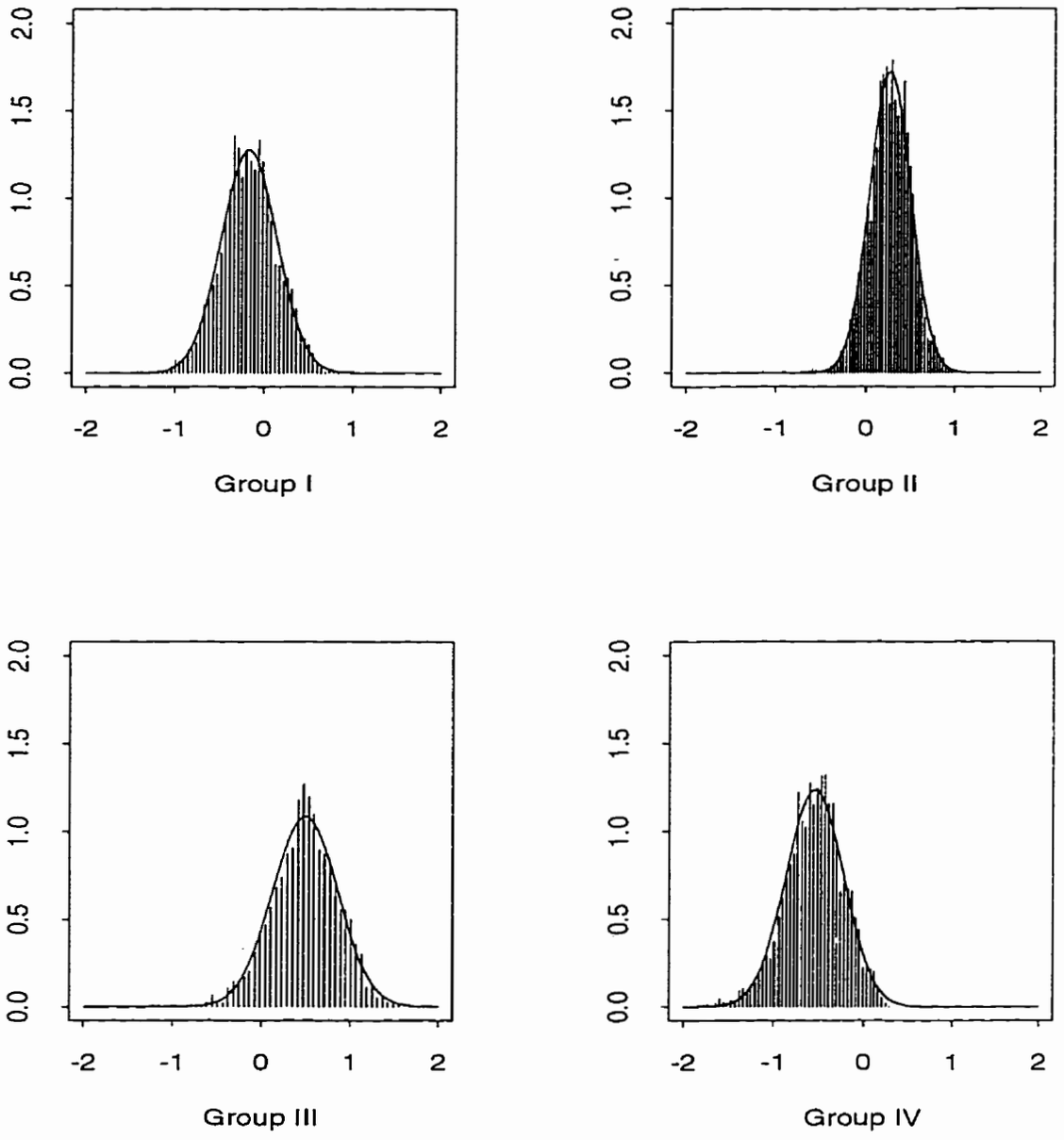


*Fig.* 4.1 Histograms of $\log \eta_k$, $1 \leq k \leq 4$ and their normal approximations.

The posteriors of $\log \eta_i$ are graphically displayed in Figure 4.1 and shown to be close to normal distributions. A brief summary of the posterior features is given in Table 4.1.

**Table 4.1**

SUMMARY OF THE BAYESIAN MANY-SAMPLE COMPARISON USING NON-INFORMATIVE PRIOR

| $i$ | $E \log \eta_i$ | $E(\log \eta_i/\eta_1)$ | $\mathrm{se}(\log \eta_i)$ | $\mathrm{se}(\log \eta_i/\eta_1)$ |
|---|---|---|---|---|
| 1 | −0.169 | — | 0.313 | — |
| 2 | 0.275 | 0.444 | 0.231 | 0.337 |
| 3 | 0.493 | 0.662 | 0.367 | 0.454 |
| 4 | −0.544 | −0.375 | 0.323 | 0.395 |

Meanwhile, we use Cox's proportional hazards regression to compare these subgroups. We set the covariate $z = (z_1, z_2, z_3)'$ and define $z = (0,0,0)'$, $(1,0,0)'$, $(0,1,0)'$, $(0,0,1)'$ for subgroups I-IV respectively. The results are summarized in Table 4.2.

**Table 4.2**

SUMMARY OF THE MANY-SAMPLE COMPARISON USING COX'S PROPORTIONAL HAZARDS REGRESSION

| | $\hat{\beta}$ | $\exp(\hat{\beta})$ | $\mathrm{se}(\hat{\beta})$ |
|---|---|---|---|
| $z_1$ | 0.425 | 1.530 | 0.341 |
| $z_2$ | 0.724 | 2.063 | 0.445 |
| $z_3$ | −0.448 | 0.639 | 0.396 |

There is only a minor difference between the Bayesian analysis and the proportional hazards regression based on Cox's partial likelihood. However, the Bayesian analysis is more flexible in terms of probability evaluation. For example, if the relative risk difference of 20% is considered significant, we may want to know the posterior probabilities $\Pr(0.8 \leq \eta_k/\eta_i \leq 1.2)$, $i \neq k$, which can be easily obtain from the sampled posterior.

## 4.4 Bayesian Prediction

Statistical prediction makes some statement about the outcome of a future experiment $E_f$ based on previous data from an informative experiment $E_i$. This kind of inference has been widely applied in clinical trials and medical survival analysis.

A Bayesian would formulate the prediction in the following way. Suppose that $E_i$ and $E_f$ are described by the same probability distribution from the class $\{F(x|\theta), \theta \in \Theta\}$, where $\Theta$ is the parameter space, and $E_i$ and $E_f$ are independent given $\theta$. Suppose, before the experiment $E_i$, a prior $\pi$ is available for $\theta$ representing our initial knowledge about the parameter. After the experiment $E_i$, the data $x$ is used to obtain the posterior $\pi(\theta|x)$, which updates our knowledge about $\theta$. The predictive distribution for the outcome $Y$ of a future experiment $E_f$ is then defined by

$$\Pr(Y \in B|x) = \int F(B|\theta)\pi(d\theta|x),$$

where $B$ is a region in the sample space and $\Pr(Y \in B|x)$ is the probability for the prediction that a future observation $Y$ will fall in $B$ based on the data $x$.

In most of the literature, $\Theta$ is of finite dimension as arises in a traditional parametric set-up. In the nonparametric case, however, $\Theta$ is infinite-dimensional. Both the prior and the posterior are stochastic processes, and characterizing their distributions is generally difficult. Ferguson (1973) introduced the well-known Dirichlet process prior which, in some cases, yields natural results. However, a major drawback of this prior is that it assigns full probability mass to the class of discrete distributions and this has posed a problem in applications. We can see this through the way that a single observation change our prediction. Suppose we have an observation $x$ and a neighbourhood $B$ of $x$. Let $\Pr(B)$ and $\Pr(B|x)$ denote the predictive probabilities based on the prior and the posterior. The local sensitivity $sen(B)$ is defined as the ratio of $\Pr(B|x) - \Pr(B)$ to $\Pr(B)$ which measures the relative change of the probability statement. We now calculate the local sensitivity for a

Dirichlet process prior with continuous intensity $\alpha$ (Ferguson, 1973) such that $\alpha(R^1) > 0$. It is easy to verify that

$$sen(B) = \frac{\alpha(R^1) - \alpha(B)}{\alpha(B)[\alpha(R^1) + 1]},$$

and $sen(B) \to \infty$ as $B$ shrinks to $x$. Therefore, the contribution of the observation has been exaggerated by the Dirichlet process prior due to its discrete nature.

When $c \to 0$, the prior introduced in section 3.1 also leads to predictive probability with zero mass allocated to intervals where no failures were observed. The data-dependent prior, however, does not have this kind of undesirable feature. Generally, the predictive probability can be expressed as

$$\Pr(T > t_k|data) = E\ \exp[-\sum_{i=1}^{k} \theta_i(\Lambda_0(t_i) - \Lambda_0(t_{i-1}))],$$

where the expectation is taken with respect to the posterior. When the conjugate prior in section 4.2 is used, we have a closed form expression

$$\Pr(T > t_k|data) = \prod_{i=1}^{k} \{1 - \frac{1}{s_i + 1 + \frac{c}{\Lambda_0(t_i) - \Lambda_0(t_{i-1})}}\}^{c+d_i}, \quad k = 1, \cdots, n. \tag{4.5}$$

We would like to point out a link between this data-dependent prior and the $A(n)$ rule, which was proposed by Hill (1968) and later generalized to censored data by Berliner and Hill (1988). The $A(n)$ rule directly specifies the predictive probabilities after the observations become available. For the uncensored data $t_1 < \cdots < t_n$, the $A(n)$ rule assigns equal probabilities $1/(n+1)$ to each of the $n+1$ open intervals $I_i = (t_i, t_{i+1})$, $i = 0, \cdots, n$ with $t_0 = 0$ and $t_{n+1} = \infty$. This prediction can be thought of as representing a robust Bayes procedure when the prior knowledge about the true distribution is extremely vague. It has been shown (Hill, 1968) that the $A(n)$ rule is an approximation to exact Bayesian procedures, though it cannot hold exactly for a continuous population.

Berliner and Hill (1988) generalized the $A(n)$ rule to censored data. Suppose that failures occur at times $t_1 < \cdots < t_n$ and there are no ties; censorings in $[t_i, t_{i+1})$ are adjusted to $t_i$.

Then the generalized $A(n)$ rule assigns to $I_i$ the predictive probability $p_i$ given by

$$p_k = \frac{1}{s_{k+1}+1} \prod_{i \le k} (1 - \frac{1}{s_i+1}). \quad k = 0, \cdots, n$$

where $s_i$ is the number of subjects at risk just before $t_i$, $i = 1, \cdots, n$ ($s_{n+1} = 0$) and an empty product is taken as one.

When $c \to 0$ the predictive probability (4.5) reduces to

$$\Pr(T > t_k | data) = \prod_{i=1}^{k} (1 - \frac{1}{s_i+1})^{d_i} \quad k = 1, \cdots, n,$$

which provides prediction for data with tied failures. If $d_i = 1$ this predictive probability coincides with $A(n)$, but Berliner and Hill (1988) suggested breaking ties arbitrarily without giving a prediction for data with tied failures.

The proportional hazards model can easily be incorporated into this prediction which yields a further generalization of $A(n)$, provided that there are no ties in the failures. Suppose the subject with covariate $z_j$ has hazard $e^{\beta' z_j} \lambda(t|\theta)$ and $s_i(\beta) = \sum_{j \in \mathcal{R}(t_i)} e^{\beta' z_j}$ where $\mathcal{R}(t_i)$ is the risk set just before $t_i$. If $\beta$ is known, by assigning a non-informative prior on the baseline hazard, the $A(n)$ can be generalized to predict the lifetime of a future subject with covariate $z$, and the corresponding predictive probability can be expressed in the same way as in (4.5) except that $s_i$ should be replaced by $s_i(\beta)\, e^{-\beta' z}$. However, $\beta$ must be estimated from the data by maximizing Cox's partial likelihood.

A pure Bayesian approach would put a prior $\pi(d\beta)$ on $\beta$ as well. For simplicity we assume $\beta$ is independent of $\theta$. Then the joint posterior would be proportional to

$$e^{\beta'(z_1 + \cdots + z_n)} \exp[-\sum_{i=1}^{n} s_i(\beta)\theta_i(\Lambda_0(t_i) - \Lambda_0(t_{i-1}))] d\theta_1 \cdots d\theta_n \pi(d\beta)$$

and the marginal posterior of $\beta$ is proportional to

$$\prod_{i=1}^{n} \frac{e^{\beta' z_i}}{s_i(\beta)} \pi(d\beta).$$

The prediction for a future subject with covariate $z$ is

$$\Pr(t_{i-1} < T < t_i | z, data) = E \frac{e^{\beta' z}}{s_i(\beta) + e^{\beta' z}} \prod_{j<i} (1 - \frac{e^{\beta' z}}{s_j(\beta) + e^{\beta' z}}), \quad i = 1, \cdots, n$$

where the expectation is with respect to the posterior of $\beta$.

Finally, we discuss the positive correlation issue in prediction. Suppose we see a patient who survives longer than $t_1$: what would be the effect on the predictive probability that next patient will survive longer that $t_2$? If $t_1 = t_2 = t$, a Bayesian would be more confident to predict that the next patient will survive longer than $t$. This is the nature of Bayesian statistics and it can be shown as follows. Let the life times of these two patients be $T_1$ and $T_2$, and suppose that for a given parameter $\theta$ they are independent. Note that if $t_1 = t_2 = t$,

$$\Pr(T_2 > t | T_1 > t) = \frac{\Pr(T_1 > t, \ T_2 > t)}{\Pr(T_1 > t)} = \frac{E_\pi [\bar{F}_\theta(t)]^2}{E_\pi \bar{F}_\theta(t)}$$

and therefore

$$\Pr(T_2 > t_2 | T_1 > t_1) - \Pr(T_2 > t_2) \geq 0, \tag{4.6}$$

where $\pi$ is the prior and $\bar{F}_\theta$ is the conditional survival function given $\theta$. But if $t_1 = t_2$, a further assumption is needed to assure the validity of (4.6). If (4.6) is valid for any $t_1$ and $t_2$, we say that $T_1$ and $T_2$ are positive dependent. In the general case, random variables $T_1, \cdots, T_n$ are said to be positive dependent if

$$\Pr(T_1 > t_1, \cdots, T_n > t_n) \geq \Pr(T_1 > t_1) \cdots \Pr(T_n > t_n) \tag{4.7}$$

for any $t_1, \cdots, t_n$.

In Bayesian statistics observations are exchangeable but not necessarily positive dependent. A simple example will illustrate this. Suppose $T_1$, $T_2$ are conditionally independent with survival function $\bar{F}_\lambda(t) = \exp(-t^\lambda)$ where $\lambda$ has a uniform prior distribution $U(1, 2)$. If $t_1 < 1 < t_2$ then (4.6) does not hold. This can be verified by the fact that

$$(b - a) \int_a^b f(x) g(x) dx \leq \int_a^b f(x) dx \int_a^b g(x) dx$$

provided one function is increasing and the other decreasing over the interval. In the nonparametric approach, however, this kind of negative correlation does not exist under a Dirichlet process or Gamma process prior, or even the squared Bessel process prior. The observations are always positively dependent, and this will be shown in the following for the Dirichlet process prior.

Suppose $F(t)$ is a Dirichlet process with intensity $\alpha(t)$, and $T_1, \cdots, T_n$ are *iid* for a given sample path of the Dirichlet process. We use mathematical induction to prove the positive dependence. Since the case $n = 1$ is trivial, we assume that (4.7) is valid for $n = k$ and proceed to show it is also correct for $n = k + 1$. Note that

$$\begin{aligned} & \Pr(T_1 > t_1, \cdots, T_k > t_k, T_{k+1} > t_{k+1}) \\ = & \int \Pr(T_1 > t_1, \cdots, T_k > t_k, T_{k+1} > t_{k+1} | T_{k+1} = y) dF_{T_{k+1}}(y) \\ = & \int_{t_{k+1}}^{\infty} \Pr(T_1 > t_1, \cdots, T_k > t_k | T_{k+1} = y) dF_{T_{k+1}}(y), \end{aligned}$$

where $F_{T_{k+1}}$ denotes the marginal distribution of $T_{k+1}$. According to Ferguson (1973) the posterior of $F(t)$ given $T_{k+1} = y$ is again a Dirichlet process with intensity $\alpha(t) + \delta(t - y)$ and, it then follows from the induction assumption that

$$\Pr(T_1 > t_1, \cdots, T_k > t_k | T_{k+1} = y) \geq \Pr(T_1 > t_1 | T_{k+1} = y) \cdots \Pr(T_k > t_k | T_{k+1} = y).$$

Without loss of generality we can assume that $t_{k+1} \geq \max_{1 \leq i \leq k} t_i$ and verify that

$$\Pr(T_i > t_i | T_{k+1} = y) \geq \Pr(T_i > t_i) \quad \text{for} \quad y \geq t_{k+1}, \ i = 1, \cdots, k,$$

and therefore the positive dependence follows.

## 4.5 Fiducial Approach

Certain restrictions apply to the fiducial inference. Let $Q(\theta, T)$ be a pivotal where $\theta$ is the parameter and $T$ is a statistic. To qualify for fiducial inference, $T$ must be a sufficient statistic

for $\theta$. Next, to guarantee the probability transform the following one-to-one correspondence must be satisfied:

(a) given $Q$ and $T$ there is a unique solution of $\theta$;

(b) given $Q$ and $\theta$ there is a unique solution of $T$.

Conditions (a) and (b) may be too restrictive in some situations. For example, (a) may be violated if for some given $Q = q$ and $T = t$ there are many values of $\theta$ satisfying $Q(\theta, t) = q$. If this is the case, all members in the set $\{\theta \mid Q(\theta, t) = q\}$ will be treated as identical because we cannot distinguish one from the other based on the information provided by $Q$. Thus, a fiducial distribution is induced on a smaller parameter space generated by grouping the "identical" parameter values as one value.

In survival analysis, there are pivotal quantities and fiducial argument can be applied. In fact, Nelson's hazard estimate is the mean of fiducial probability though he does not obtain it in that way. For simplicity we do not consider censoring at this time. Suppose the cumulative hazard function of a population is $\Lambda$, and $T_1 < \cdots < T_n$ are ordered failure times. It is well known that $(n - i + 1)[\Lambda(T_i) - \Lambda(T_{i-1})]$, $i = 1, \cdots, n$ are *iid* standard exponential variables. In other words, $Q(\Lambda, T)$ is a pivotal quantity with $\Lambda$ as the parameter. Once the data $t$ are available, we could use the fiducial step to argue that the distribution of $Q(\Lambda, t)$ is unchanged. But the parameter here is of infinite dimension and condition (a) is obviously violated. Therefore, a fiducial distribution is induced only to a projected space of cumulative hazard functions where two members $\Lambda_1$ and $\Lambda_2$ are considered the same if $\Lambda_1(t_i) = \Lambda_2(t_i)$, $i = 1, \cdots, n$.

With this we would have a finite-dimensional distribution of $\Lambda(t)$ at times $t_1 < \cdots < t_n$. This information is partial or incomplete since fiducial statements for other margins are missing. But one should not be blamed for only being able to draw inferences on such a margin with sample $(t_1, \cdots, t_n)$ at hand. The limitation is not due to the procedure of inference but to the data available. We used to add an assumption on the cumulative hazard.

such as a step function assumption. But this cannot be derived from the fiducial probability and is of course additional to the information furnished by the data.

Now, the mean of the fiducial probability for $\Lambda(t_i)$ is easily calculated as

$$E\,\Lambda(t_k) = \sum_{i=1}^{k} \frac{1}{n-i+1}$$

and so is the variance

$$\text{var}\,\Lambda(t_k) = \sum_{i=1}^{k} \frac{1}{(n-i+1)^2}.$$

Thus, like the Bayesian predictive distribution, a fiducial predictive probability can also be obtained. Let $t_0 = 0$ and $t_{n+1} = \infty$ and denote an empty product by one. For a future observation $T$, the predictive fiducial probability would be

$$\begin{aligned}
\Pr\left(t_{k-1} < T < t_k\right) &= E\left[\exp(-\Lambda(t_{k-1})) - \exp(-\Lambda(t_k))\right] \\
&= \prod_{i=1}^{k-1} \frac{n-i+1}{n-i+2} - \prod_{i=1}^{k} \frac{n-i+1}{n-i+2} \\
&= \frac{1}{n+1}. \qquad k = 1, \cdots, n+1,
\end{aligned}$$

which is exactly the value specified by the $A(n)$.

Fisher had tried to develop a coherent theory of pivotal inference but failed. In practice it is not always the case that a pivotal quantity can be found. For example, when right censoring is present, $Q$ is no longer a pivotal and the fiducial argument must be modified. Suppose that failures were observed at times $t_1 < \cdots < t_n$ and, censorings that occured in $[t_i, t_{i+1})$ are adjusted to $t_i$. Let $s_i$ denote the number of subjects at risk just before time $t_i$. We show that $Q$ is in some sense a pivotal quantity. Let us study the survival experience sequentially. From time $t_{i-1}$ to $t_i$ which is called the $i$-th period, a total of $s_i$ subjects were under observation. Let $T_1, \cdots, T_{s_i}$ be the life times of these $s_i$ subjects. At the beginning of the $i$-th period we had the information that $\min(T_1, \cdots, T_{s_i}) > t_{i-1}$ and what we observed at the end of the period is $\min(T_1, \cdots, T_{s_i})$. It is straightforward to show that, conditioning on the history before $t_{i-1}$, specifically, $s_i$ and $t_{i-1}$, the cumulative hazard for $\min(T_1, \cdots, T_{s_i})$

is $s_i[\Lambda(t) - \Lambda(t_{i-1})]$ and thus the sampling distribution for $s_i[\Lambda(t_i) - \Lambda(t_{i-1})]$ is standard exponential. If the experiment can be viewed as a sequence of independent experiments, the joint distribution of $s_i[\Lambda(t_i) - \Lambda(t_{i-1})]$, $i = 1, \cdots, n$ is independent standard exponential.

This argument is not mathematically rigorous, but the sequential view of an experiment is helpful in dealing with survival problems. For example, Cox's partial likelihood is built by sequentially identifying the subject who fails first in each risk group. The many-sample comparison (Kalbfleisch and Prentice, 1980) is carried out by analyzing a sequence of contingency tables. This approach can sometime simplify the situation and approximate the exact result.

With the fiducial probability on $s_i[\Lambda(t_i) - \Lambda(t_{i-1})]$, $i = 1, \cdots, n$ we can draw inference on the life distribution or future observation. It can be seen that the fiducial predictive probability is exactly the $A(n)$ prediction. Survival curve estimation is possible if we choose a kind of interpolation to compensate for the missing information within interval $(t_{i-1}, t_i)$. For example, right continuous step function interpolation is commonly used. The step function type of estimate is obtained by taking the mean of the fiducial probability

$$E \exp[-\Lambda(t)] = \prod_{t_i \le t} (1 - \frac{1}{s_i + 1}).$$

which is only slightly different from the Kaplan-Meier estimate. Fiducial probability intervals can be obtained by easy simulations or even by closed-form formulae. As an example, we show how to evaluate the fiducial probability $\Pr(a < \Lambda(t_k) < b)$ where $0 < a < b$. Note that, if $Y_1, \cdots, Y_n$ are independent standard exponential variables, then the linear combination $\mu_1 Y_1 + \cdots + \mu_n Y_n$ has a density function

$$\sum_{i=1}^{n} \frac{\mu_i^{n-1}}{\prod_{j \ne i} (\mu_i - \mu_j)} \frac{1}{\mu_i} \exp(-\frac{y}{\mu_i}), \quad y > 0. \tag{4.8}$$

where $\mu_i > 0$ are distinct numbers. From (4.8) we have

$$\Pr(a < \Lambda(t_k) < b) = \sum_{i=1}^{k} \frac{1}{\prod_{j \ne i} (1 - \frac{s_i}{s_j})} [\exp(-s_i a) - \exp(-s_i b)], \quad k = 1, \cdots, n.$$

which is easy to compute.

*CHAPTER 5*

# *Analysis of Binary Data*

## 5.1 Introduction

Binary data arise from experiments in which observations can be classified into two categories. For instance, in a social survey, the attitude of a subject towards a proposal may be positive or negative; in a medical experiment, a test animal may die or survive from a given dose of a poisonous drug. Statistical theory concerning binary data is extensive and involves many applications. Our primary interest here is the analysis of binary data from medical or biological experiments.

Bioassay is an experimental procedure for evaluating the biological potency of a material such as a therapeutic drug or carcinogenic substance. It has been used for many years for a wide range of purposes, and the analysis of bioassay experiments is an important part of statistics. In a typical bioassay, a dose denoted by $t$ is administered to each subject who either does or does not respond. For example, in a carcinogenicity experiment, an animal may or may not develop a tumour after a period of being regularly fed a carcinogenic chemical, and we say it responded if a tumour is detected. Suppose that, at a fixed dose $t$, a subject would respond with probability $P(t)$, a quantity of primary interest. We call $P(t)$ the response probability at dose $t$; or the dose-response relation, or the potency curve when it is viewed as a function of $t$.

Quantitative cancer risk assessment has been an important application of modern bioassay. Many of the standard methods of cancer risk assessment are based on binary data. For example, see Mantel and Bryan (1961). In this kind of approach, similar animals are exposed

to carcinogens at several different dose levels. After a fixed period, all animals are autopsied and the presence or absence of tumors in the target organ of each animal is observed. With the assumption of the functional form of the dose-response curve, the parameters can be estimated by likelihood methods. We can then estimate the VSD (virtually safe dose), a very small dose at which the risk of cancer will be less than a given probability. This often involves estimation outside the range of observations and hence is termed low-dose extrapolation.

The statistical analysis of bioassay is essentially the same as non-linear regression. With a link function $F$, which is usually a cumulative probability distribution function, the dose-response curve is expressed as $P(t) = F(\alpha + \beta \log t)$, $\beta > 0$. For example, in logistic regression $F(t) = e^t/(1 + e^t)$, and in probit regression $F$ is the cumulative probability distribution function of the standard normal distribution. The maximum likelihood method can be applied under a constraint $\beta > 0$. The disadvantage is mainly the strong model assumption on the dose-response relation, which may be inadequate in a certain range of dose levels yet fit the data quite well. For example, Van Ryzin (1980) showed that low-dose extrapolation is very sensitive to the model assumptions, that is, different choices of $F$ produce very different results.

The nonparametric view of a bioassay is that of binomial inference under order restrictions. This, and more general problems such as isotonic regression, are extensively studied by Barlow *et al* (1972). An efficient algorithm for estimation is available but convenient access to confidence limits is missing. Another obvious drawback is that the observed data are only available at the experimental levels and very little inference can be drawn about a dose other than the experimental values. Perhaps a kind of extrapolation is needed to complete the analysis.

Ramsey (1972) was one of the early works using the Bayesian method to analyze bioassay data. Antoniak (1974) also studied the same problem by introducing a mixture of Dirichlet process priors, but the implementation is difficult. A more recent work of Gelfand and Kuo

(1991) proposed a resampling approach to the Bayesian bioassay.

This chapter investigates the posterior computation problem for bioassay data using Dirichlet process priors. A nonparametric method for combining many assays is also proposed and illustrated. Finally, a possible extension to the analysis of doubly-censored data is discussed.

## 5.2 Binomial Inference Under Order Restrictions

Bioassay can be viewed as an order-restricted binomial problem. Suppose a bioassay is performed at dose levels $t_1 < \cdots < t_s$, and the dose-response curve $P(t)$ is an increasing function. There are two ways to summarize our initial knowledge. The first way is direct parameterization by setting $P(t_i) = \theta_i$. If no initial knowledge is available one might use the maximum entropy prior for $\theta = (\theta_1, \cdots, \theta_s)$ which is uniform on the simplex $0 < \theta_1 < \cdots < \theta_s < 1$. Suppose the experimental result shows $r_i$ out of $n_i$ subjects responded at dose level $t_i$. The posterior density of $\theta$ is then proportional to

$$\theta_1^{r_1}(1-\theta_1)^{n_1-r_1}\cdots\theta_s^{r_s}(1-\theta_s)^{n_s-r_s}I(0<\theta_1<\cdots<\theta_s<1) \tag{5.1}$$

where $I$ is the indicator function.

The mode of (5.1), was obtained by Ayer *et al* (1955) in a nearly closed form expression. The Gibbs sampling approach was first studied by Gelfand and Kuo (1991). The sampling scheme is quite obvious since the conditional distribution of $\theta_i$ given $\theta_{[-i]} = (\theta_1, \cdots, \theta_{i-1}, \theta_{i+1}, \cdots, \theta_s)$ is $Beta(r_i + 1, n_i - r_i + 1)$ doubly truncated at $\theta_{i-1}$ and $\theta_{i+1}$, with $\theta_0 = 0$ and $\theta_{s+1} = 1$. More generally, we may use the ordered product Beta prior with density proportional to

$$\theta_1^{r'_1}(1-\theta_1)^{n'_1-r'_1}\cdots\theta_s^{r'_s}(1-\theta_s)^{n'_s-r'_s}I(0<\theta_1<\cdots<\theta_s<1) \tag{5.2}$$

which leads to a tractable posterior. Ayer's algorithm and the Gibbs sampler as well can be applied to the posterior using (5.2) as a prior. This is numerically equivalent to modifying the observations as $r_i + r_i'$ out of $n_i + n_i'$ responded instead of the original experimental result.

The second prior for $\theta$ can be constructed by summarizing the initial knowledge on the differences $\theta_i - \theta_{i-1}$, $i = 1, \cdots, s+1$ using a Dirichlet distribution. If the prior for $\theta_i - \theta_{i-1}$, $i = 1, \cdots, s+1$ is $D(b_1, \cdots, b_{s+1})$, then the joint distribution of $\theta$ is the so-called ordered Dirichlet with density

$$\pi(\theta) \propto \prod_{i=1}^{s+1} (\theta_i - \theta_{i-1})^{b_i - 1} \; I(0 < \theta_1 < \cdots < \theta_s < 1).$$

Under the bioassay data given above the posterior density is proportional to

$$\prod_{i=1}^{s} \theta_i^{r_i} (1 - \theta_i)^{n_i - r_i} \prod_{i=1}^{s+1} (\theta_i - \theta_{i-1})^{b_i - 1} \; I(0 < \theta_1 < \cdots < \theta_s < 1). \tag{5.3}$$

At first glance, one may easily give up this approach because the sampling scheme is not obvious. In fact, the conditional distribution of $\theta_i$ given $\theta_{[-i]}$ has a density proportional to

$$\theta_i^{r_i} (1 - \theta_i)^{n_i - r_i} (\theta_i - \theta_{i-1})^{b_i - 1} (\theta_{i+1} - \theta_i)^{b_{i+1} - 1} I(\theta_{i-1} < \theta_i < \theta_{i+1}),$$

which seems unfamiliar. However, a different approach gives a quite different perspective. Note that

$$(\theta_i - \theta_{i-1})^{b_i - 1} = b_i \int_0^\infty \int_0^1 \xi_i^{-2} \eta_i^{b_i - 1} I(\eta_i < \theta_i - \theta_{i-1} < \xi_i) d\xi_i d\eta_i,$$

and thus (5.3) is a marginal density arising from the joint density proportional to

$$\prod_{i=1}^{s} \theta_i^{r_i} (1 - \theta_i)^{n_i - r_i} \prod_{i=1}^{s+1} \xi_i^{-2} \eta_i^{b_i - 1} I(\eta_i < \theta_i - \theta_{i-1} < \xi_i)$$

and obtained by integrating out $\xi = (\xi_1, \cdots, \xi_{s+1})$ and $\eta = (\eta_1, \cdots, \eta_{s+1})$. Efficient Gibbs sampling for $(\xi, \eta, \theta)$ is available. Thus, instead of sampling the posterior of $\theta$ directly, we consider the joint density of $(\xi, \eta, \theta)$ with $\xi$ and $\eta$ serving as auxiliary variables. Let the current value of these parameters be $(\theta^{(k)}, \xi^{(k)}, \eta^{(k)})$. The updating of $\theta$ must be done

sequentially to retain the order. Suppose we have updated the value of $\theta_j$, $j \leq i-1$. To update $\theta_i$ we generate

$$\theta_i^{(k+1)} \sim Beta(r_i + 1,\ n_i - r_i + 1)$$

subject to left and right truncation at

$$\max(\theta_{i-1}^{(k+1)} + \eta_i^{(k)},\ \theta_{i+1}^{(k)} - \xi_i^{(k)}) \quad \text{and} \quad \min(\theta_{i+1}^{(k)} - \eta_{i+1}^{(k)},\ \theta_{i-1}^{(k+1)} + \xi_i^{(k)})$$

where $\theta_0 = 0$ and $\theta_{s+1} = 1$ are fixed. Given $\theta$, the auxiliary variables are conditionally independent and thus can be simultaneously updated by generating *iid* variables $U_i$, $V_i \sim U(0,1)$, $1 \leq i \leq s+1$ and letting

$$\xi_i^{(k+1)} = (\theta_i^{(k+1)} - \theta_{i-1}^{(k+1)})/U_i, \qquad \eta_i^{(k+1)} = (\theta_i^{(k+1)} - \theta_{i-1}^{(k+1)})\ V_i^{1/b_i}.$$

It is obvious that the auxiliary variables have simplified the sampling scheme and updating these auxiliary variables costs very little compared to the updating of the original parameters.

Inference on a particular dose level, not necessarily an experimental level, is of practical importance. For example, in cancer risk assessment, interest centers on very low doses. The ordered binomial model, however, leaves this extrapolation problem unaddressed for the dose is not explicitly included in the model. The order restrictions only incorporate dose information in a very weak manner. To carry out an extrapolation, additional assumptions concerning the dose-response relationship are required.

## 5.3 Tolerance Distribution Approach

The tolerance of a subject is a threshold value $T$ such that when the stimulus exceeds $T$ the subject will fail and otherwise it will survive. Let us assume that the tolerance $T$ is a random variable across the population. The tolerance distribution function $P(t)$ is thus the potency curve on which our initial knowledge might be summarized using a Dirichlet process. If an initial estimate $P_0(t)$ for the tolerance distribution $P(t)$ is available with a

certain degree of confidence, then a Dirichlet process with shape parameter $P_0(t)$ and a properly chosen confidence $c$ can be used to summarize this knowledge. Let $t_1 < \cdots < t_s$ be the experimental dose levels and $\theta_i = P(t_i)$, $i = 1, \cdots, s$. Under the Dirichlet process prior, the joint distribution of $(\theta_1, \cdots, \theta_s)$ is the so-called ordered Dirichlet with density

$$\pi(\theta) \propto \prod_{i=1}^{s+1} (\theta_i - \theta_{i-1})^{b_i - 1} \; I(0 < \theta_1 < \cdots < \theta_s < 1)$$

where $b_i = c[P_0(t_i) - P_0(t_{i-1})]$, $i = 1, \cdots, s+1$ and $P_0(t_0) = 0$. Suppose the data are the same as that described in the last section, then the posterior density for $\theta$ is

$$p(\theta) \propto \prod_{i=1}^{s} \theta_i^{r_i} (1 - \theta_i)^{n_i - r_i} \prod_{i=1}^{s+1} (\theta_i - \theta_{i-1})^{b_i - 1} \; I(0 < \theta_1 < \cdots < \theta_s < 1).$$

and the auxiliary variables technique discussed in section 5.2 applies for Gibbs sampling. Gelfand and Kuo (1991) also proposed a Gibbs sampling scheme for this posterior but their method is less efficient.

Prediction and estimation for the response probability at an arbitrary dose level is also possible. Suppose $t_{i-1} < t < t_i$: then, under the posterior, the conditional distribution of $P(t)$ given $P(t_{i-1})$ and $P(t_i)$ is easily characterized. In fact,

$$Y_i(t) = \frac{P(t) - P(t_{i-1})}{P(t_i) - P(t_{i-1})}, \qquad i = 1, \cdots, s+1,$$

are $s+1$ conditionally independent Dirichlet processes with shape

$$\frac{P_0(t) - P_0(t_{i-1})}{P_0(t_i) - P_0(t_{i-1})}$$

and confidence $c[P_0(t_i) - P_0(t_{i-1})]$, which is the same as that under the prior. Therefore, data information enters $P(t)$ only through $\theta_{i-1} = P(t_{i-1})$ and $\theta_i = P(t_i)$. Once a sample of $\theta$ is available, inference on the whole dose-response curve or tolerance distribution becomes feasible. For example, the posterior for any quantile of the tolerance distribution is obtained as follows. Suppose the $q$-th quantile of $P(t)$ is $\xi_q$: then, for $t_{i-1} < t < t_i$,

$$\Pr(\xi_q \le t) \;=\; \Pr(P(t) > q)$$

$$= E \Pr(P(t) \geq q | P(t_{i-1}) = \theta_{i-1}, P(t_i) = \theta_i)$$
$$= E \Pr(Y_i(t) \geq \frac{q - \theta_{i-1}}{\theta_i - \theta_{i-1}}).$$

where $Y_i$ is the Dirichlet process defined above and the expectation is taken with respect to the posterior of $\theta$. From section 1.2 we have

$$\Pr(Y_i(t) \geq \frac{q - \theta_{i-1}}{\theta_i - \theta_{i-1}}) = 1 - B(\frac{q - \theta_{i-1}}{\theta_i - \theta_{i-1}} \mid c[P_0(t) - P_0(t_{i-1})], c[P_0(t_i) - P_0(t)])$$

where $B(x|\alpha, \beta)$ denotes the cumulative probability function of the Beta distribution with parameters $\alpha$ and $\beta$. Therefore, the posterior distribution of any quantile can be computed numerically if a sample of $\theta$ is available.

As an application, we consider the low-dose extrapolation based on the data from Mantel and Bryan (1961), reproduced in Table 5.1. After a single injection of methylcholanthine into mice at different dose levels shown below, the number of mice that developed tumours after a fixed period was recorded.

**Table 5.1**

NUMBERS OF MICE DEVELOPED TUMOURS AT DIFFERENT DOSES

| Dose | Number of tumour / Number of mice |
|---|---|
| $2.44 \times 10^{-4}$ | 0/158 |
| $9.75 \times 10^{-4}$ | 0/79 |
| $1.95 \times 10^{-3}$ | 0/38 |
| $3.9 \times 10^{-3}$ | 0/19 |
| $7.8 \times 10^{-3}$ | 3/17 |
| $1.56 \times 10^{-2}$ | 6/18 |
| $3.12 \times 10^{-2}$ | 13/20 |
| $6.25 \times 10^{-2}$ | 17/21 |
| $1.25 \times 10^{-1}$ | 21/21 |

The conventional method for analyzing this kind of data is based on logistic or probit regression. The logistic and probit are special cases of a more general approach called

generalized probit regression which allows the use of an arbitrary probability distribution function $F$ as the link function. Unfortunately, our interest in this problem lies in the left end of the dose-response curve, which is very sensitive to the choice of the link function.

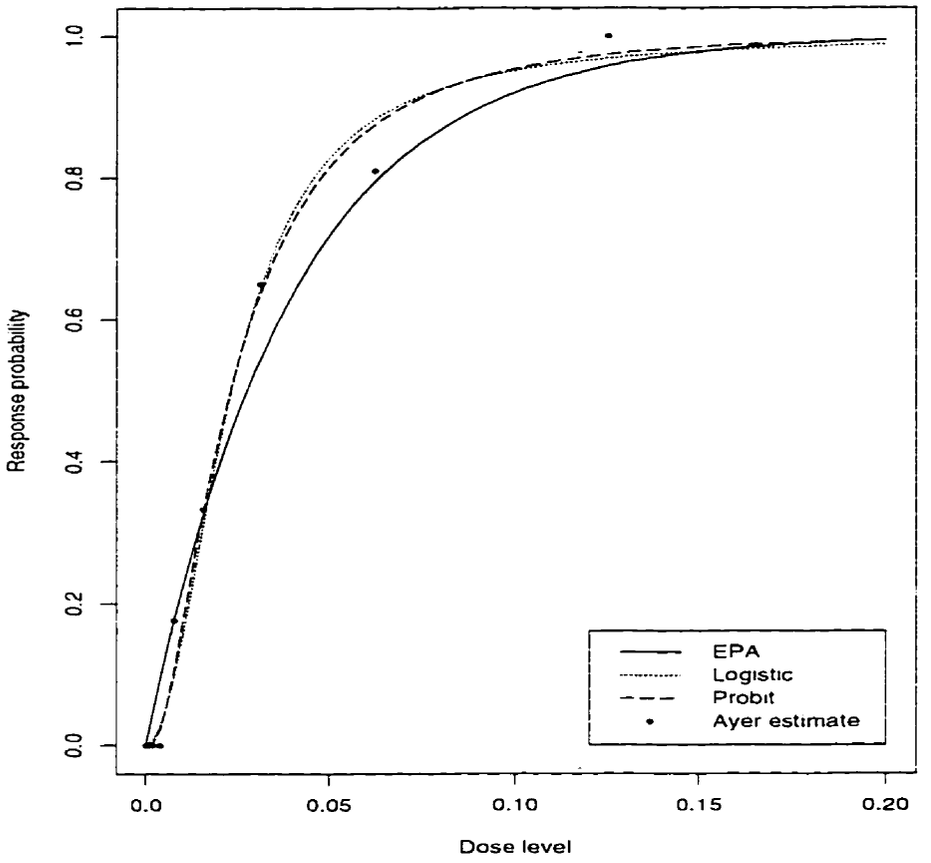


*Fig.* 5.1 Comparison of the estimated dose-response relations in the range of experimental dose levels.

We compare the estimated dose-response curves from logistic regression, probit regression and EPA (Environmental Protection Agency) method which assumes an exponential tolerance distribution. Figure 5.1 shows that all the models fit the data very well. However,

in cancer risk assessment, our primary concern is often focused on the VSD (Virtually Safe Dose), which is defined as the dose corresponding to a chance of one in a million of developing tumour. Figure 5.2 shows that, when $t$ is extremely small the tail behaviour of $F$ dominates the magnitude of VSD, and the effect of the parameters cannot compensate for the huge difference.

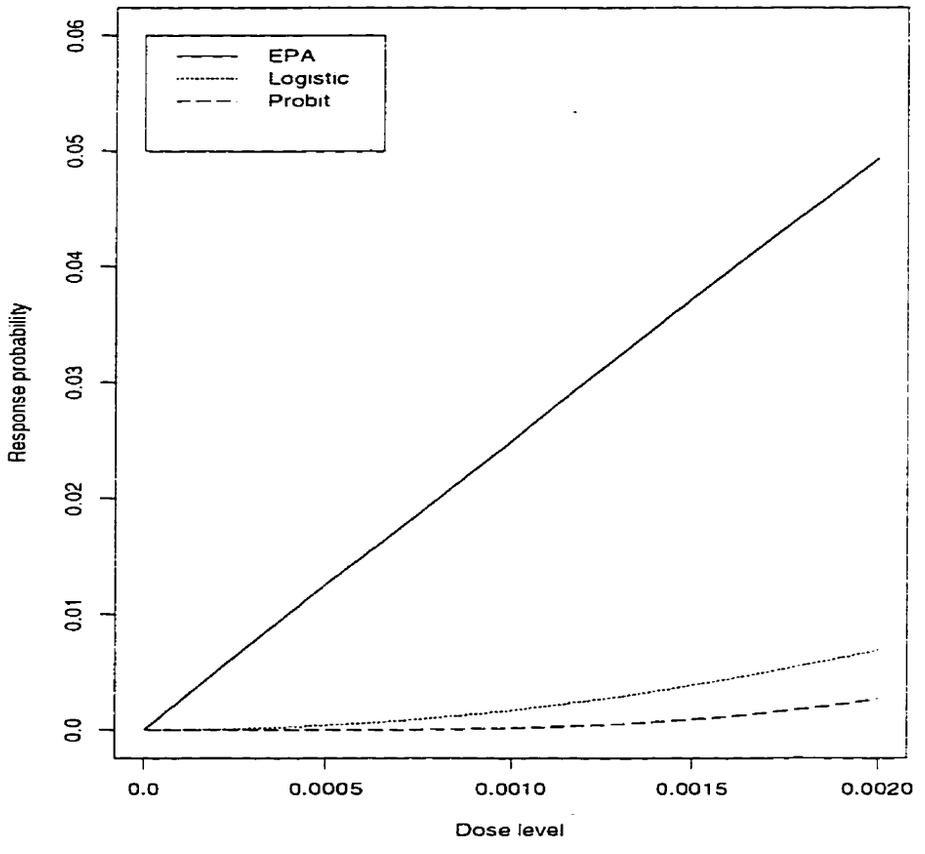


*Fig.* 5.2 Comparison of the estimated dose-response relations in the range of low dose levels.

On the other hand, the data may support many different models with very different tail behaviour. Thus, there is no way to choose a better model based on goodness-of-fit or any

other statistical rule. A more fundamental theory or insight into tumour formation is needed to give a scientifically plausible model.

One possible solution is to seek a functional form of the dose-response relation based on the mechanism of cell progression. The one-hit model (Holland and Sielken, 1993), which is derived from the assumption that cancer originates from a single cell which progresses through several irreversible changes, implies a dose-response relation

$$P(t) = 1 - \exp(-\sum_{i=0}^{k} q_i t^i) \tag{5.4}$$

where $q_i \geq 0$.

The EPA method chooses a linearized version of (5.4) for simplicity, and fits the model under the constraints $q_i \geq 0$, $i = 0, 1$ using maximum likelihood. Suppose the estimates are $(\hat{q}_0, \hat{q}_1)$. Then the VSD can be estimated by

$$\hat{VSD} = [-\log(1 - 10^{-6}) - \hat{q}_0]/\hat{q}_1.$$

For the data in Table 5.1, the EPA method gives $\hat{q}_0 = 0$, $\hat{q}_1 = 25.3$ and $\hat{VSD} = 3.95 \cdot 10^{-8}$; the fitted dose-response curve is displayed in Figure 5.1. Exact confidence limits for the VSD can be obtained by using the non-parametric Bootstrap described in Holland and Sielken (1993).

A Bayesian approach is proposed here that accommodates model uncertainty and attaches weights to all possible models by using a Dirichlet process prior. Suppose we have a favoured model, for example, the model used by EPA. We would assign a Dirichlet process prior for $P(t)$, the tolerance distribution. The shape parameter is chosen as $P_0(t) = 1 - \exp[-(q_0 + q_1 t)]$ with $q_0$ and $q_1$ unspecified, and the confidence $c$ now becomes a measure of model uncertainty. This can be viewed as a relaxed version of EPA method because we put weights on a continuous spectrum of possible models though the EPA model is at the center. In practice, several confidence levels can be chosen to see the sensitivity of the VSD on model uncertainty.

An important issue is the estimation of $q_0$ and $q_1$, for the prior depends on these unknown quantities. Once this is solved, the posterior of the tolerance distribution can be obtained by the approach we just presented. Bayesian inference on VSD is thus available since the VSD is a particular quantile of the tolerance distribution. Following the conventional empirical Bayes approach, we choose $(q_0, q_1)$ such that the average likelihood $\int L(\theta)\pi(d\theta|q_0, q_1)$ reaches its maximum. However, computing this average likelihood to generate a surface could be difficult. The amount of computation is substantially reduced if one of the parameters can be estimated from other sources. Note that if data are available from a control group, $q_0$ is easily estimated since $1 - \exp(-q_0)$ is the response probability for a subject in the control group. This is possible even when the current experiment does not have a control group. Detailed discussions are available in Grice and Ciminera (1988) on utilizing historical control data.

Our task now is only that of estimating $q_1$. We compute the average likelihood for a series of values of $q_1$ to generate a curve. For this, we first generate a large sample $\theta^{(k)}, k = 1, \cdots, N$ from the density proportional to

$$\theta_1^{r_1}(1-\theta_1)^{n_1-r_1}\cdots\theta_s^{r_s}(1-\theta_s)^{n_s-r_s}I(0<\theta_1<\cdots<\theta_s<1)$$

either by rejection sampling or Gibbs sampling, depending on the rejection rate. Then, for a fixed $q_1$,

$$\int L(\theta)\pi(d\theta|q_0, q_1) \simeq \frac{\Gamma(b_1+\cdots+b_{s+1})}{\Gamma(b_1)\cdots\Gamma(b_{s+1})}\frac{1}{N}\sum_{k=1}^{N}\prod_{i=1}^{s+1}(\theta_i^{(k)}-\theta_{i-1}^{(k)})^{b_i-1}.$$

where $b_i = c\exp(-q_0)[\exp(-q_1 t_{i-1}) - \exp(-q_1 t_i)]$. The EPA model can be fit using standard software for generalized linear models. This offers an initial range for $q_1$, within which the average likelihood will be computed and the maximum identified.

For the data in Table 5.1, we assume that the tolerance distribution is a Dirichlet process with shape $P_0(t) = 1 - \exp[-(q_0 + q_1 t)]$. Strong evidence in the current data indicates that $q_0 = 0$ though historical control data are not available here. We thus assume $P_0(t) = 1 -$

$\exp(-q_1 t)$ and the estimated $q_1$ is listed in Table 5.2: the likelihood functions are depicted in Figure 5.3. The estimated value of $q_1$ is quite stable with the increasing of model uncertainty, but the accuracy of the estimate diminishes significantly. This can be seen from Figure 5.3. The cumulative posterior probability functions of the VSD are displayed in Figure 5.4 corresponding to different degrees of model uncertainty. The posterior probability intervals for the VSD tabulated in Table 5.2 are different in that, with increasing model uncertainty, the probability has shifted to the right and become more spread out.

**Table 5.2**

BAYESIAN INFERENCE ON THE MODEL PARAMETER AND VSD

| Confidence | $\hat{q}_1$ | Posterior median of VSD | Symmetric 90% probability interval of VSD |
|---|---|---|---|
| 20 | 28.4 | $1.19 \times 10^{-4}$ | $2.10 \times 10^{-5} \sim 4.90 \times 10^{-4}$ |
| 50 | 27.6 | $6.21 \times 10^{-5}$ | $4.65 \times 10^{-6} \sim 2.38 \times 10^{-4}$ |
| 100 | 27.2 | $3.42 \times 10^{-5}$ | $2.55 \times 10^{-6} \sim 1.43 \times 10^{-4}$ |
| 200 | 25.8 | $1.86 \times 10^{-5}$ | $1.39 \times 10^{-6} \sim 7.85 \times 10^{-5}$ |
| 500 | 25.8 | $8.01 \times 10^{-6}$ | $5.99 \times 10^{-7} \sim 3.36 \times 10^{-5}$ |
| 1000 | 25.8 | $4.38 \times 10^{-6}$ | $3.28 \times 10^{-7} \sim 1.83 \times 10^{-5}$ |
| $\infty$ | 25.3 | $3.91 \times 10^{-8}$ | $3.20 \times 10^{-8} \sim 4.78 \times 10^{-8}$ * |

* Note: the bottom row is the result from EPA analysis and the probability statement is based on a nonparametric Bootstrap.

This example shows that the inference on low dose response is not only sensitive to the model selection such as the logistic, probit or EPA, but also to the model uncertainty. The EPA conservative estimate of the VSD is $3.20 \times 10^{-8}$ which is obtained by taking the lower bound from the 90% probability interval based on the one-hit model. However, if we are not completely certain about the model, say $c = 1000$, a conservative estimate of the VSD would be $3.28 \times 10^{-7}$, which is substantially larger than the EPA assessment.

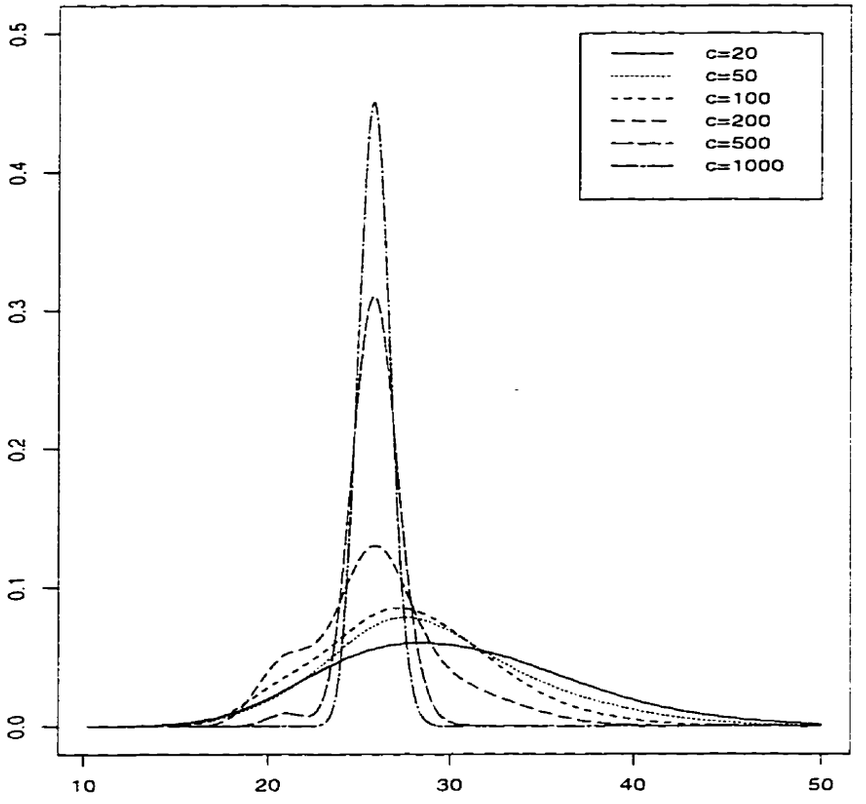


*Fig.* 5.3 Comparison of the re-scaled marginal likelihood functions for $q_1$ at different confidence levels.

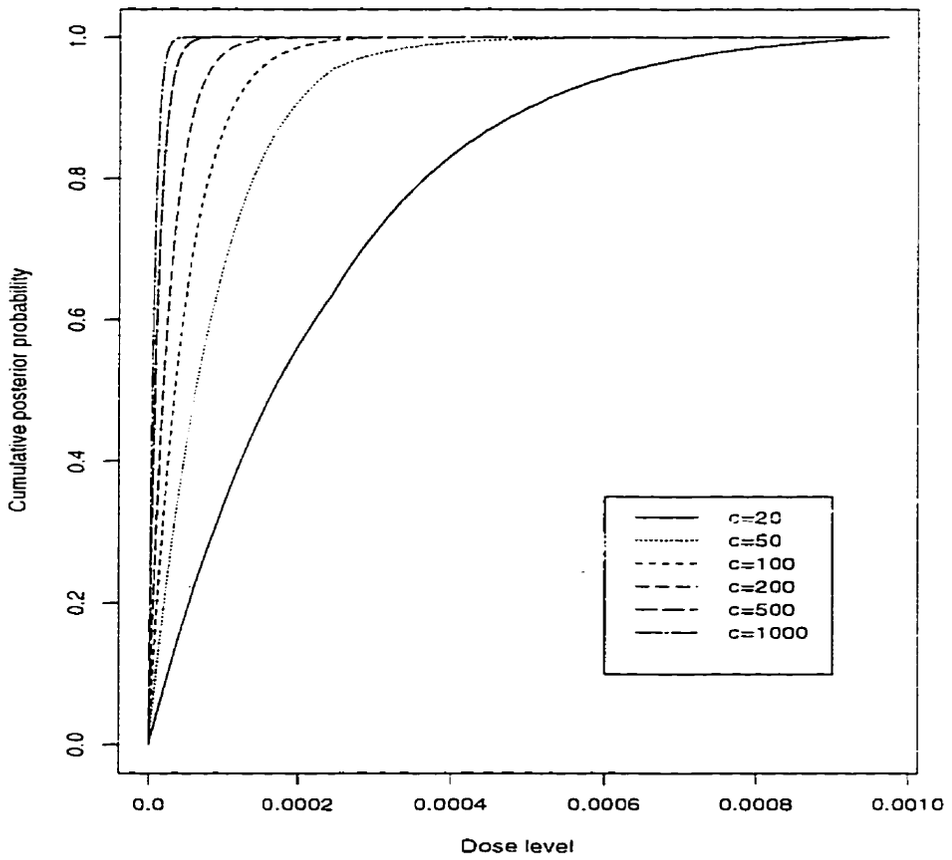


*Fig.* 5.4 Comparison of the cumulative posterior probability functions for the VSD at different confidence levels.

## 5.4 Many-Sample Problem

It is often necessary to combine many bioassay experiments since the physical limitation may prevent a laboratory from executing the single large assay needed to achieve the required precision. In such cases, the experiment may be repeated over time, or a co-operative study may involve several groups of investigators.

Using the Beta distribution to describe the variation between binomial populations is a traditional approach. In animal toxicological experiments, variation in the response from subjects is to be expected between treatments as well as between litters. It is often the case that, when litter effect is ignored, the true standard errors of estimated treatment differences will be substantially under-estimated. A two-way analysis of variance for binary data is needed when both treatment and litter effects are considered. Williams (1975) proposed a method using a Beta-binomial model to describe the extra variation between litters. Related discussions can be found in Crowder (1978), Williams (1982) and Conaway (1990). But these studies do not cover the bioassays if no specific dose-response curve is assumed.

In the Bayesian framework, hierarchical modeling is appropriate for this kind of problem. Suppose that, for testing the biological potency of a drug, $m$ bioassay experiments have been independently carried out and have produced the following results:

$$\begin{array}{cccc} r_{11}/n_{11} & r_{12}/n_{12} & \cdots & r_{1s}/n_{1s} \\ r_{21}/n_{21} & r_{22}/n_{22} & \cdots & r_{2s}/n_{2s} \\ \cdots & \cdots & \cdots & \cdots \\ r_{m1}/n_{m1} & r_{m2}/n_{m2} & \cdots & r_{ms}/n_{ms} \end{array}$$

with each row being the output of a bioassay. Due to experiment-to-experiment variation, the response probability at a fixed dose can vary across different assays. If we believe the experiment-to-experiment variation is entirely random, a plausible modeling is to assume that the response probabilities at a fixed dose in different assays are independently generated from the same population.

Let $\pi_{ij}$ be the response probability in the $i$-th assay and $j$-th dose level. We postulate an additive random effect model

$$\log \frac{\pi_{ij}}{1-\pi_{ij}} = \eta_i + \mu_j. \qquad i = 1, \cdots, m, \quad j = 1, \cdots, s. \tag{5.5}$$

where $\mu = (\mu_1, \cdots, \mu_s)$ is the dose effect and $\eta = (\eta_1, \cdots, \eta_m)$ is a random effect representing variation between experiments. We still use the non-parametric approach that no specific

dose-response relation is assumed except a natural constraint that $\mu_1 < \cdots < \mu_s$. However, even for a simple model like (5.5), an exact analysis could be technically very complicated.

When the batch sizes $n_{ij}$ are not very small, the empirical logistic transform (Cox, 1970) works quite well. This method utilizes the approximation

$$y_{ij} = \log(\frac{r_{ij} + \frac{1}{2}}{n_{ij} - r_{ij} + \frac{1}{2}}) \sim N(\log \frac{\pi_{ij}}{1 - \pi_{ij}}, v_{ij}),$$

where the term $\frac{1}{2}$ is used to guard against extreme observations, and

$$v_{ij} = \frac{(n_{ij} + 1)(n_{ij} + 2)}{n_{ij}(r_{ij} + 1)(n_{ij} - r_{ij} + 1)}$$

is the asymptotic variance. We can thus build a generalized linear model with normal error as

$$y_{ij} = \eta_i + \mu_j + \epsilon_{ij},$$

where $\eta_i \sim N(0, \sigma_\eta^2)$ and $\epsilon_{ij} \sim N(0, v_{ij})$ are all independent random variables. If $\sigma_\eta^2$ is known, the model can be reduced to

$$\bar{y}_{\cdot j} = \bar{\eta} + \mu_j + \bar{\epsilon}_{\cdot j},$$

where $\bar{\eta} \sim N(0, \sigma_\eta^2/m)$. Bayesian analysis can be simply done using a non-informative prior for $\mu$. Suppose we know very little about these parameters. A diffuse prior for $\mu$ would be the uniform distribution on the simplex $-\infty < \mu_1 < \cdots < \mu_s < \infty$. This prior can be thought of as the distribution of order statistics from a normal distribution with extremely large variance. The conditional posterior of $\mu$ given $\bar{\eta} \sim N(0, \sigma_\eta^2/m)$ is expressed as

$$\mu_j \sim N(\bar{y}_{\cdot j} + \bar{\eta}, \bar{v}_{\cdot j}/m)$$

subject to the order constraint $\mu_1 < \cdots < \mu_s$. We can sample the posterior by generating $\xi_j \sim N(\bar{y}_{\cdot j}, \bar{v}_{\cdot j}/m)$ and then test whether $\xi_j$ is ascending in $j$. If this is true we generate $\bar{\eta} \sim N(0, \sigma_\eta^2/m)$ and set $\mu_j = \bar{\eta} + \xi_j$; otherwise we return to the first step.

However, in some cases, rejection sampling could be costly. An alternative Gibbs sampling scheme is quite simple. If the current value is $(\mu^{(k)}, \bar{\eta}^{(k)})$, we update the parameters

by first generating

$$\bar{\eta}^{(k+1)} \sim N(\frac{\sum_{j=1}^{s}(\bar{y}_{\cdot j} - \mu_j^{(k)})/\bar{v}_{\cdot j}}{\sum_{j=1}^{s} 1/\bar{v}_{\cdot j}}, (\sum_{j=1}^{s} \frac{m}{\bar{v}_{\cdot j}})^{-1})$$

and then

$$\mu_j^{(k+1)} \sim N(\bar{y}_{\cdot j} - \bar{\eta}^{(k+1)}, \bar{v}_{\cdot j}/m)$$

left and right truncated at $\mu_{j-1}^{(k+1)}$ and $\mu_{j+1}^{(k)}$.

We now discuss the issue of estimating $\sigma_\eta^2$, the magnitude of the random effect. Averaging (5.5) over the second index we have

$$\bar{y}_{i\cdot} = \bar{\mu} + \eta_i + \bar{\epsilon}_{i\cdot} .$$

where the variance of $\bar{\epsilon}_{i\cdot}$ is known to be $\bar{v}_{i\cdot}/s$. There is an unbiased estimate for $\sigma_\eta^2$ given by

$$\hat{\sigma_\eta^2} = \frac{1}{m-1}\sum_{i=1}^{m}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 - \frac{1}{m-1}(1 - \frac{1}{s})\bar{v}_{\cdot\cdot}.$$

which is commonly used in the traditional analysis of variances with random effects. The main drawbacks are that it is not fully efficient and could be negative.

Another estimate can be derived using a likelihood approach. Note that the full likelihood of $(\bar{\mu}, \sigma_\eta^2)$ is readily available if $\eta_i + \bar{\epsilon}_{i\cdot}$ is taken as the error term. Then, a marginal likelihood for $L(\sigma_\eta^2)$ is obtained by integrating out the nuisance parameter $\bar{\mu}$, which is uniformly distributed over the whole real space. This gives

$$\begin{aligned} -2\log L(\sigma_\eta^2) &= \sum_{i=1}^{m}\log(\sigma_\eta^2 + \bar{v}_{i\cdot}/s) + \log(\sum_{i=1}^{m}\frac{1}{\sigma_\eta^2 + \bar{v}_{i\cdot}/s}) \\ &\quad + (\sum_{i=1}^{m}\frac{\bar{y}_{i\cdot}}{\sigma_\eta^2 + \bar{v}_{i\cdot}/s})^2 (\sum_{i=1}^{m}\frac{1}{\sigma_\eta^2 + \bar{v}_{i\cdot}/s})^{-1} - \sum_{i=1}^{m}\frac{\bar{y}_{i\cdot}^2}{\sigma_\eta^2 + \bar{v}_{i\cdot}/s}. \end{aligned}$$

**Table 5.3**

NUMBERS OF MICE SURVIVING CHALLENGE WITH BORDETELLA PERTUSSIS AFTER INOCULATION WITH STANDARD VACCINE

| Assay No. | Day/Mth. in 1970 | Dose of vaccine | | |
|---|---|---|---|---|
| | | 0.2 | 1.0 | 5.0 |
| 1 | 5/1 | 8 | 16 | 19 |
| 2 | 5/1 | 6 | 13 | 27 |
| 3 | 15/1 | 9 | 17 | 27 |
| 4 | 22/1 | 10 | 14 | 29 |
| 5 | 29/1 | 4 | 11 | 25 |
| 6 | 5/2 | 13 | 24 | 32 |
| 7 | 12/2 | 10 | 16 | 31 |
| 8 | 19/2 | 7 | 15 | 29 |
| 9 | 26/2 | 5 | 13 | 23 |
| 10 | 25/3 | 3 | 14 | 24 |
| 11 | 2/4 | 3 | 18 | 22 |
| 12 | 16/4 | 2 | 17 | 23 |
| 13 | 23/4 | 4 | 15 | 28 |
| 14 | 20/5 | 5 | 12 | 27 |
| 15 | 4/6 | 6 | 12 | 25 |
| 16 | 11/6 | 9 | 25 | 32 |
| 17 | 18/6 | 10 | 19 | 29 |
| 18 | 25/6 | 6 | 20 | 27 |
| 19 | 23/7 | 2 | 8 | 26 |
| 20 | 6/8 | 11 | 18 | 27 |
| 21 | 27/8 | 4 | 13 | 19 |
| 22 | 3/9 | 4 | 12 | 20 |
| 23 | 10/9 | 5 | 16 | 24 |
| 24 | 17/9 | 4 | 17 | 31 |
| 25 | 24/9 | 11 | 24 | 31 |
| 26 | 30/9 | 10 | 19 | 28 |
| 27 | 8/10 | 9 | 15 | 26 |
| 28 | 5/11 | 6 | 20 | 32 |
| 29 | 26/11 | 5 | 18 | 23 |
| 30 | 9/12 | 8 | 23 | 29 |
| 31 | 17/12 | 9 | 23 | 27 |
| 32 | 31/12 | 7 | 14 | 28 |

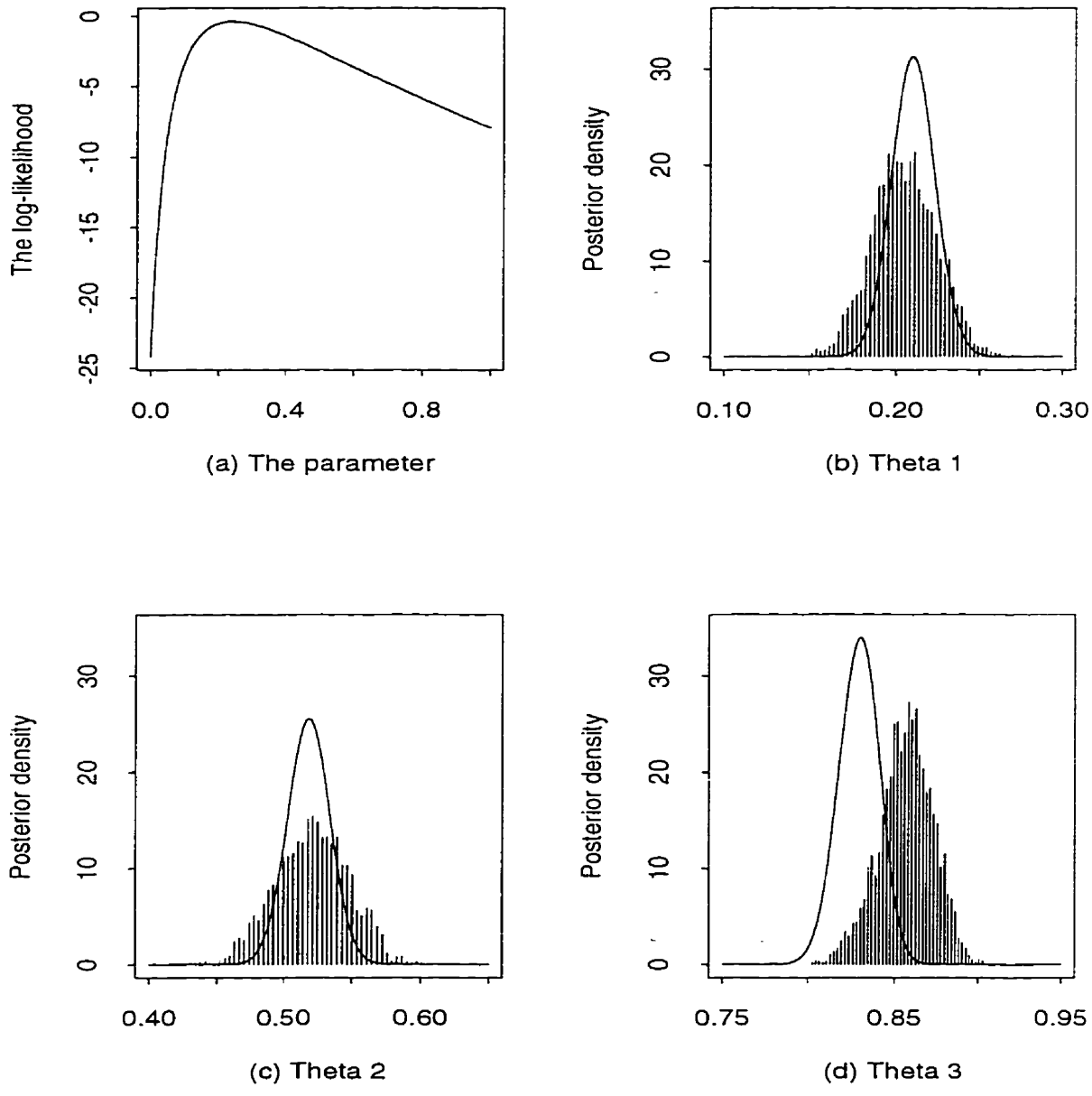


*Fig.* 5.5 (a) Log-likelihood of the variance of the random effect (b)-(d) Comparison of the two Bayesian analyses by pooled data and by a random effect model.

We illustrate our method using a dataset from Finney *et al* (1975). Table 5.3 shows the result of a series of bioassays conducted in 1970 by British National Institute for Biological Standards and Control. In each assay, the same 3 doses of standard vaccine were used with 32 mice per dose. The numbers of mice that survived the challenge with bordetella pertussis were noted.

For this particular dataset, the estimate $\hat{\sigma}_\eta^2 = 0.24$ based on the log-likelihood $\log L(\sigma_\eta^2)$ is plotted in Figure 5.5(a). The posterior density of $\theta_j = \exp(\mu_j)/(1+\exp(\mu_j))$ is depicted in Figure 5.5(b)-(d) by impulse type plots. Another Bayesian analysis can be performed by pooling the data and using the non-informative prior proposed in section 5.2: the corresponding posterior densities are represented by solid lines in Figure 5.5(b)-(d). It can be seen that the posteriors for $\theta_j$ based on the pooled data and on the random effect model are quite different in that the random effect model gives larger posterior variances to the parameters. We are thus less confident about any conclusion drawn from this posterior than that derived from a single large assay. For example, the random effect model gives posterior probability $\Pr(|\theta_2 - E\,\theta_2| \leq 0.02) = 0.536$; while the corresponding probability should be 0.796 if the data were from a single assay. Therefore, the information available is less than would be the case if all assays were homogeneous and could be combined.

It is also noted that the posterior means of $\theta_j$ from these two analyses do not agree. The random effect model gives a larger posterior mean at high dose and a smaller posterior mean at low dose than that of the pooled data analysis. This is mainly caused by the logistic transform, which is concave over $(0, \frac{1}{2})$ and convex over $(\frac{1}{2}, 1)$. At the high dose, where $r_{i3}/n_{i3}$, $i = 1, \cdots, m$ all exceed $\frac{1}{2}$, by Jensen's inequality we have

$$\bar{y}_{\cdot 3} \geq \log \frac{r_{\cdot 3}}{n_{\cdot 3} - r_{\cdot 3}} \tag{5.6}$$

for all $n_{ij}$ are the same. For the random effect model, we have $E\,\mu_3 \geq \bar{y}_{\cdot 3}$ due to left truncation. Next, since the posterior of $\mu_3$ is mostly concentrated in $(0, \infty)$, applying Jensen's

inequality to both the logistic transform and its inverse yields

$$E\,\theta_3 \geq \frac{\exp(E\,\mu_3)}{1+\exp(E\,\mu_3)} \geq \frac{\exp(\bar{y}_{\cdot 3})}{1+\exp(\bar{y}_{\cdot 3})}. \tag{5.7}$$

where the expectation is with respect to the posterior from the random effect model. Combining (5.6) and (5.7) we have

$$E\,\theta_3 \geq \frac{r_{\cdot 3}}{n_{\cdot 3}}.$$

Strictly, $r_{\cdot 3}/n_{\cdot 3}$ is less than the posterior mean of $\theta_3$ from the pooled data analysis. But due to the pooling of the data, the order-restricted Beta distribution becomes close to the Beta distribution. Thus, $r_{\cdot 3}/n_{\cdot 3}$ is close to the posterior mean of $\theta_3$ from the pooled data analysis. This explains the disagreement in posterior mean at the highest dose from the two Bayesian analyses. The lowest dose case can be explained in a similar way.

## 5.5 Further Topics

Doubly-censored failure time data can be regarded as a mix of survival and binary data. Suppose that exact failure times are observed at $t_1 < \cdots < t_n$ with $d_i$ deaths at time $t_i$ and $l_i$ and $r_i$ left- and right-censored observations respectively: exactly $s_i$ subjects are at risk just before time $t_i$. Suppose the survival function is horizontal between failure times and $\theta_i = S(t_i)$, $i = 1, \cdots, n$. Then a likelihood for $\theta = (\theta_1, \cdots, \theta_n)$ is given by

$$L(\theta) = \prod_{i=1}^{n} (\theta_{i-1} - \theta_i)^{d_i} \theta_i^{r_i} (1-\theta_i)^{l_i}. \tag{5.8}$$

from which the maximum likelihood estimate can be derived and viewed as a generalized Kaplan-Meier estimate.

A nonparametric estimate for the survival function under doubly-censored data was derived by Turnbull (1974). The estimate is obtained by an iterative procedure. First, an initial estimate of the survival function $\hat{S}(t)$ is easily obtained as the Kaplan-Meier estimate by ignoring all the left-censored observations. We use this initial estimate to modify the

data so that the left-censored observations are incorporated into the numbers of deaths and thus the Kaplan-Meier estimate can be applied. Let

$$w_{ij} = \frac{\hat{S}(t_{i-1}) - \hat{S}(t_i)}{1 - \hat{S}(t_j)}.$$

The number of deaths and the number of subjects at risk are now modified as

$$d_i' = d_i + \sum_{j \geq i} w_{ij} l_j$$

and

$$s_i' = s_i + \sum_{j \geq i} d_j'.$$

A new Kaplan-Meier estimate can be obtained and thus one iteration is completed. Turnbull (1974) also showed that this procedure converges to the maximum likelihood estimate derived from (5.8). However, a convenient method for calculating the confidence limits is missing.

A Bayesian approach to this problem can be carried out under the same parameterization. In the following we only work with the posterior under non-informative priors, yielding results comparable to Turnbull's estimate. Note that the parameters $\theta_i$ are automatically ordered and a non-informative prior would be the uniform distribution on the simplex $\theta$. $1 \geq \theta_1 > \cdots > \theta_n \geq 0$. The posterior density is found to be proportional to the product of the likelihood (5.8) and the indicator of the simplex. However, direct Gibbs sampling from this posterior is difficult.

A possible solution involves introducing auxiliary variables. For convenience we assume $\theta_0 = 1$ and $\theta_{n+1} = 0$. It is easily verified that the posterior density for $\theta$ can be expressed as

$$p(\theta) \propto \prod_{i=1}^{n} (\theta_{i-1} - \theta_i)^{d_i} \theta_i^{r_i} (1 - \theta_i)^{l_i} I(1 \geq \theta_1 > \cdots > \theta_n \geq 0)$$

which is a marginal density of

$$p(\theta, \eta) \propto \prod_{i=1}^{n} \eta_i^{d_i - 1} I(\eta_i < \theta_{i-1} - \theta_i) \theta_i^{r_i} (1 - \theta_i)^{l_i} I(1 \geq \theta_1 > \cdots > \theta_n \geq 0)$$

by integrating out the auxiliary variable $\eta$. As we previously demonstrated in section 5.2, Gibbs sampling for $(\eta, \theta)$ can be easily carried out. Let $(\theta^{(k)}, \eta^{(k)})$ denote the current value of the parameters. Suppose we have updated the value of $\theta_j$, $j \leq i-1$. To update $\theta_i$, we generate

$$\theta_i^{(k+1)} \sim Beta(r_i + 1, l_i + 1)$$

subject to left and right truncation at $\theta_{i+1}^{(k)} + \eta_{i+1}^{(k)}$ and $\theta_{i-1}^{(k+1)} - \eta_i^{(k)}$ respectively. Given $\theta$, $\eta_i$, $i = 1, \cdots, n+1$ are conditionally independent and thus can be updated by generating $iid$ variables $U_i \sim U(0,1)$, $i = 1, \cdots, n+1$ and letting

$$\eta_i^{(k+1)} = (\theta_{i-1}^{(k+1)} - \theta_i^{(k+1)}) U_i^{1/d_i}, \quad i = 1, \cdots, n+1.$$

This is even simpler than the sampling scheme proposed in section 5.2 for bioassay data.

A full Bayesian analysis for doubly-censored data using Ferguson's Dirichlet process prior is also very easy. Note that the Dirichlet process prior is conjugate when all the observations are complete. Suppose the prior is a Dirichlet process with shape $F_0$ and confidence $c$; after incorporating the complete observations that $d_i$ subjects failed at $t_i$, the posterior is again a Dirichlet process with shape

$$F_0'(t) = \frac{cF_0(t) + \sum_{i=1}^n d_i I(t \geq t_i)}{c + \sum_{i=1}^n d_i}$$

and confidence $c' = c + \sum_{i=1}^n d_i$. This posterior is used as a prior now to further extract the information in the left- and right-censored observations. We would not repeat the story told in section 5.3 where a Dirichlet process prior is used to bioassay data. The pure left- and right-censored data can be viewed as the result of an assay conducted at dose levels $t_1 < \cdots < t_n$ with tolerance distribution $F(t)$. The final posterior can be obtained exactly in the way as we previously outlined in section 5.3.

*CHAPTER 6*

# *Discussion and Summary*

## 6.1 Discussion

Constructing the smoothed prior is motivated by some criticisms of the Dirichlet process for its discrete sample paths. After this long journey struggling with computation, however, we need a serious re-thinking of the role of smoothing, both pro and con. Generally, smoothing is not central in applications yet it remains an active research topic in mathematical statistics. Smoothing techniques are occasionally useful in signal or image processing. But inference is another story. In most situations, statistical analysis for one-sample problems is descriptive in nature. Therefore, a rough image such as Kaplan-Meier estimate will do the job and smoothing seems superfluous.

Then what happens if we completely ignore the smoothness or parameter correlation? No obvious problem if we think the data is discrete, which is natural from a practical viewpoint. However, modeling continuous failure time data without considering any dependence in the parameters (Cornfield and Detre, 1977) is inappropriate; a correct treatment of the pointwise independent hazard function (Kalbfleisch and MacKay, 1978) should be a Gamma process approach to discrete data.

The effect of smoothing in Bayesian analysis can be explained in the light of information. Suppose that there are two priors $\pi_1$ and $\pi_2$, available for the parameter $\theta = (\theta_1, \cdots, \theta_n)$ that, under both priors each $\theta_i$ has the same marginal distribution. Therefore, marginally, the same amount of uncertainty is assumed in each parameter. But the total uncertainty in $\theta$, probably measured by entropy, depends on the joint distribution. Suppose further that, $\pi_1$

specifies independence between components of $\theta$ while $\pi_2$ incorporates a kind of dependence. Then it can be argued or even mathematically proved that $\pi_2$ is more informative and contains less uncertainty, or in other words, $\pi_1$ is more unbiased and less informative; for a given sample, $\pi_2$ leads to a narrower posterior interval for each $\theta_i$ than $\pi_1$. Thus, the degree of smoothness or dependence, which is purely a subjective assumption in most cases, actually eliminates some uncertainty in the prior, and failing to consider this may exaggerate the total uncertainty in parameters.

The implementation of a Bayesian analysis under the smoothed prior is more efficient than it appears. But further improvement can be made provided a more efficient generator of Bessel random numbers is available. In addition to the aspect of posterior computation, some intrinsic disadvantages of the nonparametric approach should also be noted. Visual exploration of the posterior is not very straightforward due to the high dimension of the parameter, and the cost of storing the posterior may not be desirable.

Analysis of various types of data is another aspect in Bayesian nonparametric statistics. This direction of research could be more application-oriented because it considers more features in the data or the experiment rather than in the prior.

## 6.2 Topics for Future Study

(i) *General smoothing structure*

For simplicity we have assumed a Markov structure on the parameters. A more general consideration would relax this condition and turn to create more symmetric relations.

(ii) *Gamma process prior for bioassay*

The Gamma process prior, or generally, the neutral to the right process prior can also be assigned to the tolerance distribution for bioassay problems. For convenience, we still use the terminology of survival analysis. Suppose we have initial information on the cumulative

hazard. Then, a Gamma process prior can be assigned to the tolerance distribution as we first described in section 1.3.

We suppose $-\log(1-P(t))$ is a gamma process with shape $\Lambda_0$ and confidence $c$, where $P(t)$ is the dose-response relation. Let $P(t_i) = 1 - \exp[-(\theta_1 + \cdots + \theta_i)]$ with $t_1 < \cdots < t_s$ being experimental levels.

If the experimental result shows $r_i$ out of $n_i$ subjects responded at dose level $t_i$, then the posterior density of $\theta = (\theta_1, \cdots, \theta_s)$ is proportional to

$$\prod_{i=1}^{s} \theta_i^{a_i - 1} \exp(-b_i\theta_i)\{1 - \exp[-(\theta_1 + \cdots + \theta_i)]\}^{r_i},$$

where $a_i = c[\Lambda_0(t_i) - \Lambda_0(t_{i-1})]$ and $b_i = c + \sum_{j=1}^{i}(n_j - r_j)$. Again, we can introduce some auxiliary variables to give a simple Gibbs sampler.

(iii) *Doubly-censored data.*

To continue the discussion broached in section 5.5, some applications and numerical examples will be considered. Theoretical issues will also be investigated, for example, the application of neutral to the right process prior to the analysis of doubly-censored data. The feasibility of this approach is already seen from our previous discussion.

(iv) *Doubly-truncated data.*

Kalbfleisch and Lawless (1992) discussed the arising of truncated data from field reliability studies and presented some statistical methods for nonparametric estimation. However, a full statistical analysis for doubly-truncated data is still a problem. We would propose a Bayesian nonparametric approach to double-truncated data using non-informative priors.

Regression and comparison between samples are practically more important, and will be studied in the future.

# APPENDIX

I. An inverse Laplace transform

LEMMA *The inverse Laplace transform of the function*

$$F(s) = \frac{1}{(Cs+D)^{\nu+1}} \exp(-\frac{As+B}{Cs+D})$$

*is*

$$f(y) = \frac{1}{C} e^{-(A+Dy)/C} (\frac{y}{\Delta})^{\nu/2} I_\nu(\frac{2\sqrt{\Delta y}}{C}),$$

*where* $\nu > -1$, $A$, , $B$, , $C$ *and* $D$ *are positive constants, and* $\Delta = AD - BC > 0$.

II. Proof for (2.17)

We now consider the evaluation of

$$U_\lambda(x,y) = E_x\,(\exp[-\lambda \int_0^1 \frac{\xi(t)}{(t+p)^2} dt] \mid \xi(1) = y).$$

First, we calculate $E_x \exp[-\int_0^1 \xi(t) d\mu(t)]$ with $d\mu = [s\delta_1 + \lambda(t+p)^{-2}]dt$ where $s > 0$ and $\delta_1$ is the *Dirac* function. It is straightforward to verify that equations (2.10) and (2.11) in this case are equivalent to the boundary problem

$$y'' - \frac{2\lambda}{(t+p)^2} y = 0,$$
$$y(0) = 1, \quad y'(1) + 2sy(1) = 0.$$

The two independent solutions of this are $(t+p)^{(1\pm\sqrt{8\lambda+1})/2}$, thus

$$y = c_1(t+p)^{(1+\sqrt{8\lambda+1})/2} + c_2(t+p)^{(1-\sqrt{8\lambda+1})/2},$$

where $c_1$ and $c_2$ are determined by the boundary conditions. After some algebra, we find that

$$y(1) = \sqrt{8\lambda+1}\sqrt{\frac{p+1}{p}} \frac{1}{Cs+D}$$

and

$$\frac{x}{2} y'(0) = -\frac{As+B}{Cs+D},$$

where

$$A = \frac{x(p+1)}{p}[\frac{\sqrt{8\lambda+1}-1}{2}(\frac{p+1}{p})^{\sqrt{8\lambda+1}/2} + \frac{\sqrt{8\lambda+1}+1}{2}(\frac{p}{p+1})^{\sqrt{8\lambda+1}/2}],$$
$$B = \frac{x\lambda}{p}[(\frac{p+1}{p})^{\sqrt{8\lambda+1}/2} - (\frac{p}{p+1})^{\sqrt{8\lambda+1}/2}],$$
$$C = 2(p+1)[(\frac{p+1}{p})^{\sqrt{8\lambda+1}/2} - (\frac{p}{p+1})^{\sqrt{8\lambda+1}/2}],$$
$$D = \frac{\sqrt{8\lambda+1}+1}{2}(\frac{p+1}{p})^{\sqrt{8\lambda+1}/2} + \frac{\sqrt{8\lambda+1}-1}{2}(\frac{p+1}{p})^{\sqrt{8\lambda+1}/2}.$$

Let $e^{\beta} = (\sqrt{8\lambda+1}+1)/\sqrt{8\lambda}$ and $e^{2\gamma} = (p+1)/p$. We can verify that

$$A = 2x\sqrt{2\lambda}e^{2\gamma}\cosh(\beta - \sqrt{8\lambda+1}\,\gamma),$$
$$B = \frac{2x\lambda}{p}\sinh\sqrt{8\lambda+1}\,\gamma,$$
$$C = 4(p+1)\sinh\sqrt{8\lambda+1}\,\gamma,$$
$$D = 2\sqrt{2\lambda}\cosh(\beta + \sqrt{8\lambda+1}\,\gamma),$$

and

$$\Delta = AD - BC = 8x\lambda e^{2\gamma}\cosh^2\beta.$$

Therefore,

$$E_x \exp[-\lambda\int_0^1 \frac{\xi(t)}{(t+p)^2}dt - s\xi(1)]$$
$$= (8\lambda+1)^{(\nu+1)/2}e^{\gamma(\nu+1)}\frac{1}{(Cs+D)^{\nu+1}}\exp(-\frac{As+B}{Cs+D}),$$

and this is equivalent to

$$\int_0^{\infty} e^{-sy}U_\lambda(x,y)q(1,x,y)dy = (8\lambda+1)^{(\nu+1)/2}e^{\gamma(\nu+1)}\frac{1}{(Cs+D)^{\nu+1}}\exp(-\frac{As+B}{Cs+D}).$$

Taking the inverse Laplace transform with respect to $s$ we obtain

$$U_\lambda(x,y) = 2(8\lambda+1)^{(\nu+1)/2}e^{\gamma(\nu+1)}\frac{1}{C}e^{-(A+Dy)/C+(x+y)/2}(\frac{x}{\Delta})^{\nu/2}I_\nu(\frac{2\sqrt{\Delta y}}{C})I_\nu^{-1}(\sqrt{xy}).$$

and it then follows that,

$$U_\lambda(x,y) = \frac{\sqrt{8\lambda+1}\sinh\gamma}{\sinh\sqrt{8\lambda+1}\,\gamma}\exp[\frac{xe^{\gamma}+ye^{-\gamma}}{2}\sinh\gamma(\coth\gamma$$
$$- \sqrt{8\lambda+1}\coth\sqrt{8\lambda+1}\,\gamma)]I_\nu(\frac{\sqrt{xy(8\lambda+1)}\sinh\gamma}{\sinh\sqrt{8\lambda+1}\,\gamma})I_\nu^{-1}(\sqrt{xy}),$$

where $e^{2\gamma} = (p+1)/p$.

# REFERENCES


Amos, D. E.(1974) Computation of modified Bessel functions and their ratios. *Math. Comp.*, **28**. 239-251.

Antoniak, C. E.(1974) Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.*, **2**, 1152-1174.

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., and Silverman, E.(1955) An empirical distribution function for sampling with incomplete information. *Ann. Math. Statist.*, **26**. 641-647.

Berliner, L. M. and Hill, B. M.(1988) Bayesian nonparametric survival analysis. *J. Amer. Statist. Assoc.*, **83**, 772-779.

Besag, J. and Green, P. J.(1993) Spatial statistics and Bayesian computation. *J. R. Statist. Soc. B*, **55**, 25-37.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M. and Brunk, H. D.(1972) *Statistical inference under order restrictions.* John Wiley, London.

Cameron, R. H. and Martin, W. T.(1944) Transformations of Wiener integrals under translations. *Ann. Math.*, **45**. 386-396.

Conaway, M. R.(1990) A random effects model for binary data. *Biometrics*, **46**, 317-328.

Cornfield, J. and Detre, C.(1977) Bayesian life table analysis. *J. R. Statist. Soc. B*, **39**. 264-296.

Cox, D. R.(1970) *The analysis of binary data.* Methuen, London.

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

Cox, D. R.(1975) Partial likelihood. *Biometrika*, **62**, 269-276.

Crowder, M. J.(1978) Beta-binomial ANOVA for proportions. *Appl. Statist.*, **27**, 34-37.

Dawid, A. P., Stone, N. and Zidek, J. V.(1973) Marginalization paradoxes in Bayesain and structural inference. *J. R. Statist. Soc. B*, **35**, 189-233.

Doksum, K. A.(1974) Tailfree and neutral random probabilities and their posterior distributions. *Ann. Prob.*, **2**, 183-201.

Dykstra, R. L. and Laud, P.(1981) A Bayesian nonparametric approach to reliability. *Ann. Statist.*, **9**, 2, 356-367.

Escobar, M. D.(1994). Estimating normal means with a Dirichlet process prior. *J. Amer. Statist. Assoc.*, **89**, 268-277.

Escobar, M. D. and West, M.(1995) Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.*, **90**, 577-588.

Feller, W.(1966). *An introduction to probability theory and its applications*, Vol. II. John Wiley and Sons, New York.

Ferguson, T. S.(1973) A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, **1**, 209-230.

Finney, D. J., Holt, L. B. and Sheffield, F.(1975) Repeated estimations of an immunological response curve. *Journal of Biological Standardization*, **3**, 1-10.

Fisher, R. A.(1935) The fiducial argument in statistical inference. *Ann. of Eugenics*, **6**, 391-398.

Gel'fand, I. M. and Yaglom, A. M.(1960) Integration in functional spaces and its applications in quantum physics. *J. Math. Phys.*, **1**, 48-69.

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

Gelfand, A. E. and Kuo, L.(1991) Nonparametric Bayesian bioassay including ordered polytomous response. *Biometrika*, **78**, 657-666.

Geman, S. and Geman, D.(1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. PAMI*, **6**, 721-741.

Grice, H. C. and Ciminera, J. L.(1988) *Carcinogenicity*. Springer-Verlag, New York.

Hill, B. M.(1968) Posterior distribution of percentiles: Bayes theorem for sampling from a finite population. *J. Amer. Statist. Assoc.*, **63**, 677-691.

Holland, C. D. and Sielken, R. L. Jr. (1993) *Quantitative cancer modeling and risk assessment*. Prentice Hall, New Jersey.

Jaynes, E. T.(1968) Prior probabilities. *IEEE Trans. SSC*, **4**, 227-291.

Jefferys, H.(1967) *Theory of probability*. Clarendon Press, Oxford.

Kalbfleisch, J. D.(1978) Nonparametric Bayesian analysis of survival time data. *J. R. Statist. Soc. B*, **40**, 214-221.

Kalbfleisch, J. D. and Lawless, J. F.(1992) Some useful statistical methods for truncated data. *Journal of Quality Tech.*, **24**, 145-152.

Kalbfleisch, J. D. and MacKay, R. L.(1978) Remarks on a paper by Cornfield and Detre. *J. R. Statist. Soc. B*, **40**, 175-177.

Kalbfleisch, J. D. and Prentice R. L.(1980) *The statistical analysis of failure time data*. John Wiley, New York.

Kimeldorf, G. S. and Wahba, G.(1970) A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Statist.*, **41**, 495-502.

Lenk, P. J.(1988) The logistic normal distribution for Bayesian, nonparametric predictive densities. *J. Amer. Statist. Assoc.*, **83**, 509-516.

Lenk, P. J.(1991) Towards a practicable Bayesian nonparametric density estimator. *Biometrika*, **78**, 531-543.

Leonard, T.(1978) Density estimation, stochastic processes and prior information. *J. R. Statist. Soc. B*, **40**, 113-146.

Lindley, D. V.(1961) The use of prior probability distributions in statistical inference and decisions. *Proc. 4th Berkeley Symp.*, **1**, 436-468.

Lo, A. Y.(1984) On a class of Bayesian nonparametric estimates. *Ann. Statist.*, **12**, 351-357.

Mantel, N. and Bryan, W. R.(1961) Safety testing of carcinogenic agents. *Journal of the National Cancer Institute*, **27**, 455-470.

Mardia, K. V.(1972) *Statistics of directional data.* Academic Press, London.

Moran, P. A. P.(1959) *The theory of Storage.* Methuen, London.

Nelson, W.(1972) Theory and applications of hazard plotting for censored failure data. *Technometrics*, **14**, 945-966.

Novick, M. R.(1969) Multiparameter Bayesian indifference procedures. *J. R. Statist. Soc. B.* **31**. 29-64.

Pitman, J. and Yor, M.(1982). A decomposition of Bessel bridges. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **59**, 425-457.

Prentice, R. L. (1974). A log-gamma model and its maximum likelihood estimation. *Biometrics*. **61**, 539-544.

Ramsey, F. L.(1972) A Bayesian approach to bio-assay. *Biometrics*, **28**, 841-858.

Smith, A. F. M. and Roberts, G. O.(1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. R. Statist. Soc. B*, **55**, 3-23.

Spain, B. and Smith, M. G.(1970). *Functions of mathematical physics.* Van Nostrand Reinhold Company, London.

Susarla, V. and Van Ryzin, J.(1976) Nonparametric Bayesian estimation of survival curves from incomplete observations. *J. Amer. Statist. Assoc.*, **71**, 896-902.

Tierney, L.(1994). Markov chains for exploring posterior distributions. *Ann. Statist.*, **22**, 1701-1762.

Titchmarsh, E. C.(1939). *The theory of functions.* Oxford Univ. Press, London.

Turnbull, B. W.(1974) Nonparametric estimation of a survivorship function with doubly censored data. *J. Amer. Statist. Assoc.*, **69**, 169-173.

Van Ryzin, J.(1980) Quantitative risk assessment. *J. Occup. Medicine*, **22**, 321-326.

Wahba, G.(1978) Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. R. Statist. Soc. B*, **40**, 364-372.

Wilks, S. S.(1962) *Mathematical statistics.* Wiley, New York.

Williams, D. A.(1975) The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics*, **31**, 949-952.

Williams, D. A.(1982) Extra binomial variation in logistic regression models. *Appl. Statist.*, **31**, 144-148.