Bayesian Computational Methods in Biomedical Research

Hedibert F. Lopes, University of Chicago*
Peter Müller, University of Texas†
Nalini Ravishanker, University of Connecticut‡

in *Computational Methods in Biomedical Research*
edited by R. Khattree and D. N. Naik. Marcel
Dekker/Taylor & Francis, pages 211-59.

1 Introduction

This article gives a survey of Bayesian techniques useful for biomedical applications. Given the extensive use of Bayesian methods, especially with the recent advent of MCMC, we can never do exhaustive justice. Nevertheless, we have made an attempt to present different Bayesian applications from different viewpoints and differing levels of complexity. We start with a brief introduction to the Bayesian paradigm in Section 2, and give some basic formulas. In Section 3, we review conjugate Bayesian analysis in both the static and dynamic inferential frameworks, and give references to some biomedical applications. The conjugate Bayesian approach is often insufficient for handling complex problems that arise in several applications. The advent of sampling based Bayesian methods has opened the door to carrying out inference in a variety of settings. They must of course be used with care, and with sufficient understanding of the underlying stochastics. In Section 4, we present details on the algorithms most commonly used in Bayesian computing and provide exhaustive references. Sections 5-8 show illustrations of Bayesian computing in biomedical applications that are of current interest.

2 A General Framework for Bayesian Modeling

Assume an investigator is interested in understanding the relationship between cholesterol levels and coronary heart disease. Both classical and Bayesian statistics start by describing the relative likelihood of possible observed outcomes as a probability model. This probability model is known as the sampling model or likelihood, and is usually indexed by some unknown parameters. For example, the parameters could be the odds of developing coronary heart disease at different levels of cholesterol. Bayesian inference describes uncertainty about the unknown parameters by a second probability model. This probability model on the parameters is known as the prior distribution. Together, the sampling model and the prior probability model describe a joint probability

*University of Chicago Graduate School of Business, 5807 South Woodlawn Avenue, Chicago, Illinois 60637. *E-mail:* hlopes@ChicagoGSB.edu.

†Department of Biostatistics, The University of Texas M. D. Anderson Cancer Center, 1515 Holcombe Boulevard, Box 447, Houston, Texas 77030-4009. *E-mail:* pm@odin.mdacc.tmc.edu.

‡Department of Statistics, University of Connecticut, CLAS 333, 215 Glenbrook Road, Storrs, Connecticut 06269. *E-mail:* nalini.ravishanker@uconn.edu.

model on the data and parameters. In contrast, classical statistics proceeds without assuming a probability model for the parameters. The prior probability model describes uncertainty on the parameters before observing any data. After observing data, the prior distribution is updated using rules of probability calculus (Bayes' rule). The updated probability distribution on the parameters is known as the posterior distribution, and contains all relevant information on the unknown parameters. From a Bayesian perspective, all statistical inference can be deduced from the posterior distribution by reporting appropriate summaries. In the rest of this section, we review the formal rules of probability calculus that are used to carry out inference, as well as model adequacy, model selection and prediction.

2.1 Discrete Case

Suppose $A_1, \ldots, A_K$ are K disjoint sets and suppose $\pi_i = P(A_i)$ is the prior probability assigned to this event, $0 \leq \pi_i \leq 1$, $\sum_{i=1}^{K} \pi_i = 1$. Consider n observable events $B_1, \ldots, B_n$. Let $p(B_j \mid A_i)$ denote the relative likelihood of the events B_j under the events A_i (sampling model). The conditional probability of A_i given the observed events is from Bayes' theorem

$$P(A_i|B_j) = \frac{P(B_j|A_i)P(A_i)}{P(B_j)} \tag{1}$$

where $P(B_j)$ is the marginal probability of observing B_j and is

$$P(B_j) = \sum_{i=1}^{K} P(B_j|A_i)P(A_i). \tag{2}$$

We often write this posterior probability as $P(A_i|B_j) \propto P(B_j|A_i)P(A_i)$. In this discrete case, the notion of Bayesian learning (updating) is described by

$$P(A_i|B_j, B_j^*) \propto P(B_j^*, B_j|A_i)P(A_i) = P(B_j^*|B_j, A_i)P(A_i|B_j). \tag{3}$$

Example 1 *A simple biomedical application discussed in Gelman et al. (2004) and Sorensen and Gianola (2002) deals with obtaining the probability that a woman XYZ is a carrier of the gene causing the genetic disease hemophilia. Double recessive women (aa) and men who carry the a allele in the X-chromosome manifest the disease. If a woman is a carrier, she will transmit the a allele with probability 0.5, and ignoring mutation, will not transmit the disease if she is not a carrier. Suppose a women is not hemophilic, her father and mother are unaffected, but her brother is hemophilic (their mother must be a carrier of a). We wish to determine the probability that XYZ is a carrier. Suppose A_1 and A_2 respectively denote the events that XYZ is a carrier, and that she is not a carrier, with (prior) probabilities $P(A_1) = P(A_2) = 0.5$. In terms of a discrete random variable θ, with $\theta = 1$ denoting she is a carrier and $\theta = 0$ indicating she is not, the prior distribution on θ is $P(\theta = 1) = P(\theta = 0) = 0.5$. The prior odds in favor of the mother not being a carrier is $P(\theta = 0)/P(\theta = 1) = 1$. Suppose information is also provided that neither of the two sons of XYZ has the disease. For $i = 1, 2$, let Y_i be a random variable assuming value 1 if the ith son has the disease and value 0 if he does not. Assuming Y_1 and Y_2 are independent conditional on θ, we have that given $\theta = 1$,*

$$P(Y_1 = 0, Y_2 = 0|\theta = 1) = P(Y_1 = 0|\theta = 1)P(Y_2 = 0|\theta = 1) = 0.5 \times 0.5 = 0.25.$$

Also, given $\theta = 0$,

$$P(Y_1 = 0, Y_2 = 0|\theta = 0) = P(Y_1 = 0|\theta = 1)P(Y_2 = 0|\theta = 1) = 1 \times 1 = 1.$$

The posterior distribution of θ can be written for $j = 0, 1$ as

$$P(\theta = j|Y_1 = 0, Y_2 = 0) = \frac{P(\theta = j)P(Y_1 = 0, Y_2 = 0|\theta = j)}{\sum_{i=0}^{1} P(\theta = i)P(Y_1 = 0, Y_2 = 0|\theta = i)}.$$

So that given that neither son is affected, the probability that XYZ is a carrier is 0.2 and that she is not a carrier is 0.8. The posterior odds in favor of XYZ not being a carrier of hemophilia is

$$P(\theta = 0|Y_1 = 0, Y_2 = 0)/P(\theta = 1|Y_1 = 0, Y_2 = 0) = 4.$$

2.2 Continuous Case

Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ be a k-dimensional vector of unknown parameters ($k \geq 1$), and suppose that *a priori* beliefs about $\boldsymbol{\theta}$ are given in terms of the probability density function (pdf) $\pi(\boldsymbol{\theta})$ (prior). Let $\mathbf{y} = (y_1, \ldots, y_n)$ denote an n-dimensional observation vector whose probability distribution depends on $\boldsymbol{\theta}$ and is written as $p(\mathbf{y}|\boldsymbol{\theta})$ (sampling model). Both $\boldsymbol{\theta}$ and $\mathbf{y}$ are assumed to be continuous-valued. To make probability statements about $\boldsymbol{\theta}$ given $\mathbf{y}$, the posterior density using Bayes' theorem is defined as (Berger 1985)

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{\pi(\boldsymbol{\theta}, \mathbf{y})}{m(\mathbf{y})} = \frac{\pi(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})}{m(\mathbf{y})} \tag{4}$$

where $m(\mathbf{y}) = \int \pi(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta}$ is the marginal density of $\mathbf{y}$ and does not depend on $\boldsymbol{\theta}$. Also called the predictive distribution of $\mathbf{y}$, $m(\mathbf{y})$ admits the marginal likelihood identity $m(\mathbf{y}) = \pi(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})/\pi(\boldsymbol{\theta}|\mathbf{y})$, and plays a useful role in Bayesian decision theory, empirical Bayes methods, and model selection. Recall that the likelihood function $l(\boldsymbol{\theta}|\mathbf{y}) = p(\mathbf{y}|\boldsymbol{\theta})$, regarded as a function of $\boldsymbol{\theta}$. We think of the prior-posterior relationship as *Posterior* $\propto$ *Prior* $\times$ *Likelihood*, and write

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \pi(\boldsymbol{\theta})l(\boldsymbol{\theta}|\mathbf{y}). \tag{5}$$

It is often convenient to work with the logarithm of the likelihood, $L(\boldsymbol{\theta}|\mathbf{y})$. Bayesian inference involves moving from a prior distribution on $\boldsymbol{\theta}$ before observing $\mathbf{y}$ to a posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y})$ for $\boldsymbol{\theta}$, and in general, consists of obtaining and interpreting $\pi(\boldsymbol{\theta}|\mathbf{y})$ via plots (contour and scatter), numerical summaries of posterior location and dispersion (mean, mode, quantiles, standard deviation, interquartile range), credible intervals (also called highest posterior density (HPD) regions), and hypotheses tests (see Lee 1997; Congdon 2003; Gelman *et al.* 2004). The sequential use of Bayes' theorem is instructive. Given an initial set of observations $\mathbf{y}$, and a posterior density (5), suppose we have a second set of observations $\mathbf{z}$ distributed independently of $\mathbf{y}$, it can be shown that the posterior $\pi(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z})$ is obtained from using $\pi(\boldsymbol{\theta}|\mathbf{y})$ as the prior for $\mathbf{z}$, i.e.,

$$\pi(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z}) \propto \pi(\boldsymbol{\theta}|\mathbf{y})l(\boldsymbol{\theta}|\mathbf{z}). \tag{6}$$

Example 2 *Lopes, Müller and Rosner (2003) consider hematologic, i.e., blood count, data from a cancer chemotherapy trial. For each patient in the trial we record white blood cell count over time as the patient undergoes the first cycle of a chemotherapy treatment. Patients are treated at different doses of the chemotherapy agent(s). The main concern is inference about the number of days that the patient is exposed to a dangerously low white blood cell count. We proceed with a parametric model for the white blood cell profile over time. In words, we assume initially a constant baseline count, followed by a sudden drop when chemotherapy is initiated, and finally a slow S-shaped recovery back to baseline after the chemotherapy. The profile is indexed by a 7-dimensional vector of random effects (see Section 3.3 below) that parameterize a non-linear regression curve that reflects these features. Let $\boldsymbol{\theta}_i$ denote this 7-dimensional vector for patient i. Let $f(t; \boldsymbol{\theta}_i)$ denote the value at time t*

for the profile indexed by $\boldsymbol{\theta}_i$. Let $y_{ij}, j = 1, \ldots, n_i$ denote the observed blood counts for patient i on (known) days t_{ij}. We assume a non-linear regression with normal residuals

$$y_{ij} = f(t_{ij};\, \boldsymbol{\theta}_i) + \epsilon_{ij}, \tag{7}$$

with a normal distributed residual error, $\epsilon_{ij} \sim N(0, \sigma^2)$. For simplicity we assume that the residual variance σ^2 is known. Model (7) defines the sampling model. The model is completed with a prior probability model $\pi(\boldsymbol{\theta}_i)$. In words, the prior reflects the judgment of likely initial white blood counts, the extent of the drop during chemotherapy, and the typical speed of recovery. Let $\mathbf{y}_i = (y_{ij},\ j = 1, \ldots, n_i)$ denote the observed blood counts for patient i. Using Bayes theorem we update the prior $\pi(\boldsymbol{\theta}_i)$ to the posterior $\pi(\boldsymbol{\theta}_i|\mathbf{y}_i)$. Instead of reporting the 7-dimensional posterior distribution, inference is usually reported by posterior summaries for relevant functions of the parameters. For example, in this application an important summary are the number of days that the patient has white blood cell count below a critical threshold. Let $f(\boldsymbol{\theta}_i)$ denote this summary. We plot $\pi(f(\boldsymbol{\theta}_i)|\mathbf{y}_i)$. In this short description we only discussed inference for one patient, i. The full model includes submodels for each patient, $i = 1, \ldots, n$, linked by a common prior $\pi(\boldsymbol{\theta}_i)$. The larger model is referred to as a hierarchical model (see Section 3.1). The prior $\pi(\boldsymbol{\theta}_i)$ is also known as the random effects distribution. It is usually indexed with additional unknown (hyper-)parameters $\boldsymbol{\phi}$ and might include a regression on patient specific covariates $\mathbf{x}_i$, in summary $\pi(\boldsymbol{\theta}_i|\boldsymbol{\phi}, \mathbf{x}_i)$. The covariate vector $\mathbf{x}_i$ includes the treatment dose for patient i. The model is completed with a prior probability model $\pi(\boldsymbol{\phi})$ for the hyperparameter ϕ. In summary, the full hierarchical model is

$$\begin{aligned} \textit{likelihood:} \quad & y_{ij} = f(t_{ij};\, \boldsymbol{\theta}_i) + \epsilon_{ij},\ i = 1, \ldots, n, j = 1, \ldots, n_i \\ \textit{prior:} \quad & \boldsymbol{\theta}_i \sim p(\boldsymbol{\theta}_i \mid \boldsymbol{\phi}, \mathbf{x}_i), \quad \boldsymbol{\phi} \sim \pi(\boldsymbol{\phi}) \end{aligned}$$

In the context of this hierarchical model, inference of particular interest is the posterior predictive distribution $\pi(\boldsymbol{\theta}_{n+1}|\mathbf{y}_1, \ldots, \mathbf{y}_n, \mathbf{x}_{n+1})$. This distribution is used to answer questions of the type: "What is the maximum dose that can be given and still bound the probability of more than 4 days below the critical lower threshold at less than 5%?".

2.3 Prior and Posterior Distributions

A prior distribution represents an assumption about the nature of the parameter $\boldsymbol{\theta}$, and clearly has an impact on posterior inference. Early Bayesian analyses dealt with conjugate priors, and this is explored further in Section 3. A prior density $\pi(\boldsymbol{\theta})$ is said to be proper if it does not depend on the data and integrates to 1. Bayesian inference is often subject to a criticism that posterior inference might be affected by choice of a subjective, injudicious prior, especially if the sample size is small or moderate. Considerable effort at defining an objective prior, whose contribution relative to that of the data is small, is often made. An extensive literature exists on approaches for specifying objective or noninformative priors (Berger and Bernardo 1992; Bernardo and Smith 1994). The uniform prior, $\pi(\theta) = 1/(b-a)$ for $\theta \in (a, b)$, is the most commonly used noninformative (vague) prior. Note that a uniform prior for a continuous parameter θ on $(-\infty, \infty)$ is improper, i.e., the integral of the pdf is not finite. While it is generally acceptable to use an improper prior, care must be exercised in applications to verify that the resulting posterior is proper. A class of improper priors proposed by Jeffreys (1961) is based on using Fisher's information measure via $\pi(\boldsymbol{\theta}) \propto |I(\boldsymbol{\theta})|^{1/2}$, where $I(\boldsymbol{\theta}) = \partial \log p(\mathbf{y}|\boldsymbol{\theta})/\partial \boldsymbol{\theta}$.

Example 3 *When $y_1, \ldots, y_n$ are independent and identically distributed (iid) $N(\mu, \sigma^2)$ with known σ^2 and $\pi(\mu) = c$, for some constant c, then it is easily verified that the posterior of μ is $\pi(\mu|\mathbf{y}) \propto \exp\{-n(2\sigma^2)^{-1}(\mu - \overline{y})^2\}$ which integrates to $\sqrt{2\pi\sigma^2/n}$ and is proper. The Jeffreys' prior is $\pi(\mu) \propto \sqrt{n/\sigma^2} = c$.*

See Bernardo (1979) for a discussion of reference priors, Sivia (1996) for examples of maximum entropy priors, and Robert (1996) and Congdon (2003) for discrete mixtures of parametric densities and Dirichlet process (DPP) priors with applications for smoothing health outcomes (Clayton and Kaldor 1987) and modeling sudden infant death syndrome (SIDS) death counts (Symons, Grimson and Yuan 1983). Prior elicitation (Kass and Wasserman 1996) continues to be an active area of research.

Posterior inference will be robust if it is not seriously affected by the choice of the model (likelihood, prior, loss function) assumptions and is insensitive to inputs into the analysis (Kadane 1984; Berger, Insua and Ruggeri 2000); see Sivaganesan (2000) for a detailed review of global robustness based on measures such as the Kullback-Leibler distance. Robustness measures are closely related to the class of priors used for analysis. Typically, a class of priors is chosen by first specifying a single prior p_0 and then choosing a suitable neighborhood Λ to reflect our uncertainty about p_0, such as ϵ-contamination classes, density bounded classes, density ratio classes, etc. For instance, an ϵ-contamination class is $\Lambda = \{p : p = (1-\epsilon)p_0 + \epsilon q; q \in Q\}$, where Q is a set of probability distributions that are possibly deviations from p_0. Kass, Tierney and Kadane (1989) has described approximate methods to assess sensitivity while Gustafson (1996) has discussed an 'informal' sensitivity analysis to compare inference on a finite set of alternative priors.

2.4 Predictive Distribution

The predictive distribution of $\mathbf{y}$ accounts both for the uncertainty about $\boldsymbol{\theta}$ and the residual uncertainty about $\mathbf{y}$ given $\boldsymbol{\theta}$, and as such, enables us to check model (prior, likelihood, loss function) assumptions. Predictive inference about an unknown observable $\tilde{y}$ is described via the posterior predictive distribution

$$p(\widetilde{y}|\mathbf{y}) = \int p(\widetilde{y}, \boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} = \int p(\widetilde{y}|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}, \tag{8}$$

where the second identity assumes that $\widetilde{y}$ and $\mathbf{y}$ are independent conditional on $\boldsymbol{\theta}$.

Example 4 *Suppose* $y_1, \ldots, y_n$ *iid* $N(\theta, \sigma^2)$, $\theta|\sigma^2 \sim N(\theta_0, \lambda_0^{-1}\sigma^2)$ *and* $\sigma^2 \sim$ *Inv-*$\chi^2(\alpha_0, \sigma_0^2)$, *i.e.* $\sigma^{-2} \sim \chi^2(\alpha_0, \sigma_0^2)$,

$$\begin{aligned} \pi(\theta|\sigma^2) &= (2\pi\lambda_0^{-1}\sigma^2)^{-1/2} \exp\left\{-0.5\lambda_0(\theta-\theta_0)^2/\sigma^2\right\} \\ \pi(\sigma^2) &= (0.5\alpha_0)^{0.5\alpha_0}\Gamma^{-1}(0.5\alpha_0)\sigma_0^{\alpha_0}\sigma^{-(\alpha_0+2)} \exp\{-0.5\alpha_0\sigma_0^2/\sigma^2\}, \end{aligned}$$

so that the joint prior $\pi(\theta, \sigma^2)$ *is given by their product as N-Inv-*$\chi^2(\theta_0, \lambda_0^{-1}\sigma^2; \alpha_0, \sigma_0^2)$ *and the joint posterior is N-Inv-*$\chi^2(\theta_1, \lambda_1^{-1}\sigma_1^2; \alpha_1, \sigma_1^2)$ *where* $\theta_1 = (\lambda_0+n)^{-1}[\lambda_0\theta_0 + n\overline{y}]$, $\lambda_1 = \lambda_0 + n$, $\alpha_1 = \alpha_0 + n$ *and*

$$\sigma_1^2 = \alpha_1^{-1}[\alpha_0\sigma_0^2 + (n-1)s^2 + (\lambda_0+n)^{-1}\lambda_0 n(\overline{y}-\theta_0)^2].$$

Suppose a noninformative prior specification $\pi(\theta, \sigma^2) \propto (\sigma^2)^{-1}$ *is used, the joint posterior specification is given by* $\pi(\theta|\sigma^2, \mathbf{y})$ *is* $N(\overline{y}, n^{-1}\sigma^2)$ *and* $\pi(\sigma^2|\mathbf{y})$ *is Inv-*$\chi^2(n-1, s^2)$.

Gelman *et al.* (2004) describe an application of a normal hierarchical model to a meta-analysis whose goal is to make combined inference from data on mortality after myocardial infarction in 22 clinical trials, each consisting of two groups of heart attack subjects randomly allocated to receive or not receive beta-blockers; see Rubin (1989) for more details on Bayesian meta-analysis.

2.5 Model Determination

Model determination consists of model checking (for adequate models) and model selection (for best model); see Gamerman and Lopes (2006, Chapter 7) and Gelman *et al.* (2004). Classical and Bayesian model choice methods would involve comparison of measures of fit to the current fitted data or cross-validatory fit to out-of-sample data. Formal Bayesian model assessment is based on the marginal likelihoods from J models M_j ($j = 1, \ldots, J$) with (i) parameter vector $\boldsymbol{\theta}_j$ whose prior density is p_j, and (ii) with prior model probability $P(M_j)$, with $\sum_j P(M_j) = 1$. Given data, the posterior model probability and the probability of the data conditional on the model (Gelfand and Ghosh 1994) are respectively

$$\begin{aligned} P(M_j|\mathbf{y}) &= P(M_j)\frac{\int l(\boldsymbol{\theta}_j|\mathbf{y})\pi(\boldsymbol{\theta}_j)d\boldsymbol{\theta}_j}{\sum_{l=1}^{J}[P(M_l)\int l(\boldsymbol{\theta}_l|\mathbf{y})\pi(\boldsymbol{\theta}_l)d\boldsymbol{\theta}_l} \\ P(\mathbf{y}|M_j) &= m_j(\mathbf{y}) = \int l(\boldsymbol{\theta}_j|\mathbf{y})\pi(\boldsymbol{\theta}_j)d\boldsymbol{\theta}_j \end{aligned} \tag{9}$$

The Bayes factor for two distinct models M_1 and M_2 is the ratio of the marginal likelihoods $m_1(\mathbf{y})$ and $m_2(\mathbf{y})$, i.e.,

$$\frac{P(\mathbf{y}|M_1)}{P(\mathbf{y}|M_2)} = \frac{P(M_1|\mathbf{y})}{P(M_2|\mathbf{y})} \times \frac{P(M_2)}{P(M_1)} \tag{10}$$

see Kass and Raftery (1995) and Pauler, Wakefield and Kass (1999) for an application to variance component models. For applications with improper priors, Bayes factors cannot be defined and several other model selection criteria have been proposed, such as the pseudo Bayes factor (Geisser 1975), intrinsic Bayes factor (Berger and Pericchi 1993), etc. Model averaging is another option in finding the best inference; see Hoeting *et al.* (1999) for a review. These methods are most effective with the sampling based approach to Bayesian inference described in Section 4.

2.6 Hypothesis Testing

The Bayesian approach to hypothesis testing of a simple $H_0 : \boldsymbol{\theta} \in \Theta_0 = \{\theta_0\}$ versus simple $H_1 : \boldsymbol{\theta} \in \Theta_1 = \{\theta_1\}$, where $\Theta = \Theta_0 \cup \Theta_1$, is more straightforward than the classical approach; it consists of making a decision based on the magnitudes of the posterior probabilities $P(\theta \in \Theta_0|\mathbf{y})$ and $P(\theta \in \Theta_1|\mathbf{y})$, or using the Bayes factor in favor of H_0 versus H_1 as

$$BF = \frac{\pi_1 P(\theta \in \Theta_0|\mathbf{y})}{\pi_0 P(\theta \in \Theta_1|\mathbf{y})}$$

where $\pi_0 = P(\theta \in \Theta_0)$ and $\pi_1 = P(\theta \in \Theta_1)$ are the prior probabilities. The Bayesian p-value is defined as the probability the probability that the replicated data could be more extreme than the observed data, as measured by the test statistic T:

$$\begin{aligned} p_B &= P(T(\mathbf{y}^{\text{rep}}, \boldsymbol{\theta}) \geq T(\mathbf{y}, \boldsymbol{\theta})|\mathbf{y}) \\ &= \int\int I_{T(\mathbf{y}^{\text{rep}},\boldsymbol{\theta})\geq T(\mathbf{y},\boldsymbol{\theta})} p(\mathbf{y}^{\text{rep}}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\mathbf{y}^{\text{rep}}d\boldsymbol{\theta} \end{aligned} \tag{11}$$

Bayesian decision analysis involves optimization over decisions in addition to averaging over uncertainties (Berger 1985). An example on medical screening is given in Gelman *et al.* (2004), Chapter 22.

3 Conjugate or Classical Bayesian Modeling

Conjugate Bayesian analysis was widely prevalent until the advent of an efficient and feasible computing framework to handle more complicated applications. A class $\mathcal{P}$ of prior distributions for $\boldsymbol{\theta}$ is naturally conjugate for a class of sampling distributions $\mathcal{F}$ if $\mathcal{P}$ is the set of all densities with the same functional form as the likelihood, and if for all densities $p(.|\boldsymbol{\theta}) \in \mathcal{F}$ and all priors $p(.) \in \mathcal{P}$, the posterior $p(\boldsymbol{\theta}|\mathbf{y})$ belongs to $\mathcal{P}$. It is well known that sampling distributions belonging to an exponential family have natural conjugate prior distributions. Specifically, suppose the sampling distribution for $\mathbf{y}$ and the prior distribution for the parameter $\boldsymbol{\theta}$ have the forms

$$\begin{aligned} p(\mathbf{y}|\boldsymbol{\theta}) &\propto g(\boldsymbol{\theta})^n \exp[\boldsymbol{\phi}(\boldsymbol{\theta})'\mathbf{t}(\mathbf{y})], \\ \pi(\boldsymbol{\theta}) &\propto g(\boldsymbol{\theta})^\eta \exp[\boldsymbol{\phi}(\boldsymbol{\theta})'\boldsymbol{\nu}] \end{aligned} \tag{12}$$

where $\mathbf{t}(\mathbf{y})$ is a sufficient statistic for $\boldsymbol{\theta}$, the posterior density for $\boldsymbol{\theta}$ has the form

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto g(\boldsymbol{\theta})^{(\eta+n)} \exp[\boldsymbol{\phi}(\boldsymbol{\theta})'(\boldsymbol{\nu} + \mathbf{t}(\mathbf{y})] \tag{13}$$

Although resulting computations are simple and often available analytically in closed forms, it has been shown that exponential families are in general the only classes of sampling distributions that have natural conjugate priors.

There are several applications in biomedical areas. A useful model in epidemiology for the study of incidence of diseases is the Poisson model. Suppose $y_1, \ldots, y_n$ is a random sample from a Poisson(μ) distribution, so that $l(\mu|\mathbf{y}) \propto \mu^{\sum y_i} \exp(-n\mu)$, which is in the exponential family. Suppose $\pi(\mu) \sim$ Gamma(α, β) (with shape α and scale β), then the posterior also belongs to the same family. Nonconjuacy is preferable, or necessary to handle most complicated problems that arise in practice. Further, analytical results may not be available and we must use simulation methods, as described in Section 4. Static and dynamic linear modeling offer a versatile class of models that may be handled using simple computational approaches under standard distributional assumptions.

3.1 Linear Modeling

Since the discussion of Bayesian inference for the linear model with a single-stage hierarchical prior structure (Lindley and Smith 1972), great strides have been made in using Bayesian techniques for hierarchical linear, generalized linear and nonlinear mixed modeling. Their two-stage hierarchical normal linear model supposes

$$\begin{aligned} \mathbf{y}|\boldsymbol{\theta}_1 \sim N(\mathbf{A}_1\boldsymbol{\theta}_1, \mathbf{C}_1), \boldsymbol{\theta}_1|\boldsymbol{\theta}_2 \sim N(\mathbf{A}_2\boldsymbol{\theta}_2, \mathbf{C}_2) \text{ and} \\ \boldsymbol{\theta}_2|\boldsymbol{\theta}_3 \sim N(\mathbf{A}_3\boldsymbol{\theta}_3, \mathbf{C}_3), \end{aligned}$$

where additionally, $\boldsymbol{\theta}_3$ is a known k_3-dimensional vector, and $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{C}_1, \mathbf{C}_2$ and $\mathbf{C}_3$ are known positive definite matrices of appropriate dimensions. The posterior distribution of $\boldsymbol{\theta}_1$ given $\mathbf{y}$ is then $N(\mathbf{Dd}, \mathbf{D})$ where

$$\begin{aligned} \mathbf{D}^{-1} &= \mathbf{A}_1'\mathbf{C}_1^{-1}\mathbf{A}_1 + [\mathbf{C}_2 + \mathbf{A}_2\mathbf{C}_3\mathbf{A}_2']^{-1} \\ \mathbf{d} &= \mathbf{A}_1'\mathbf{C}_1^{-1}\mathbf{y} + [\mathbf{C}_2 + \mathbf{A}_2\mathbf{C}_3\mathbf{A}_2']^{-1}\mathbf{A}_2\mathbf{A}_3\boldsymbol{\theta}_3. \end{aligned} \tag{14}$$

The mean of the posterior distribution is seen to be a weighted average of the least squares estimate $(\mathbf{A}_1'\mathbf{C}_1^{-1}\mathbf{A}_1)^{-1}\mathbf{A}_1'\mathbf{C}_1^{-1}\mathbf{y}$ of $\boldsymbol{\theta}_1$ and its prior mean $\mathbf{A}_2\mathbf{A}_3\boldsymbol{\theta}_3$, and is a point estimate of $\boldsymbol{\theta}_1$. The three-stage hierarchy can be extended to several stages. Smith (1973) examined the Bayesian linear model in more detail and studied inferential properties. There is an extensive literature on the application of these methods to linear regression and analysis of designed experiments

in biomedical research. Classical Bayesian inference for univariate linear regression to model responses $\mathbf{y} = (y_1, \ldots, y_n)$ as a function of an observed predictor matrix $\mathbf{X}$ stems from

$$\mathbf{y}|\boldsymbol{\beta}, \sigma^2, \mathbf{X} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}); \quad \pi(\boldsymbol{\beta}, \sigma^2|\mathbf{X}) \propto \sigma^{-2} \tag{15}$$

where the noninformative prior specification is adequate in situations when the number of cases n is large relative to the number of predictors p. Posterior inference follows from

$$\begin{aligned} \boldsymbol{\beta}|\sigma^2, \mathbf{y} &\sim N(\widehat{\boldsymbol{\beta}}, \sigma^2\mathbf{V}_{\boldsymbol{\beta}}); \quad \sigma^2|\mathbf{y} \sim \text{Inv} - \chi^2(n-p, s^2), \text{ where} \\ \widehat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}; \quad \mathbf{V}_{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}; \quad s^2 = (n-p)^{-1}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \end{aligned}$$

The conjugate family of prior distributions is the normal-Inv-χ^2 shown in Section 2. Jeffreys' prior is $\pi(\boldsymbol{\beta}, \sigma^2) \propto \sqrt{|\mathbf{I}(\boldsymbol{\beta}, \sigma^2)|} \propto (\sigma^2)^{-(p+2)/2}$. An extension to normal multivariate regression with q-variate independently distributed responses $\mathbf{y}_1, \ldots, \mathbf{y}_n$ is straightforward:

$$\mathbf{y}_i|\boldsymbol{B}, \Sigma, \mathbf{x}_i \sim N_q(\mathbf{x}_i'\boldsymbol{B}, \Sigma), i = 1, \cdots, n; \quad \pi(\boldsymbol{\beta}, \sigma^2|\mathbf{X}) \propto \sigma^{-2} \tag{16}$$

A noninformative prior specification is the multivariate Jeffreys prior $\pi(\boldsymbol{\beta}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(q+1)/2}$, and the corresponding posterior distribution is $\boldsymbol{\Sigma}|\mathbf{y} \sim \text{Inv-Wishart}_{n-1}(\mathbf{S})$ and $\boldsymbol{\beta}|\boldsymbol{\Sigma}, \mathbf{y} \sim N(\overline{\mathbf{y}}, n^{-1}\boldsymbol{\Sigma})$. The conjugate prior family for $(\mathbf{B}, \boldsymbol{\Sigma})$ is normal-Inv-Wishart$(\boldsymbol{B}_0, \boldsymbol{\Sigma}/\lambda_0, \nu_0, \Lambda_0^{-1})$ distribution. Several applications exist in the literature. Buonaccorsi and Gatsonis (1988) discussed inference for ratios of coefficients in the linear model with applications to slope-ratio bioassay, comparison of the mean effects of two soporific drugs and a drug bioequivalence problem in a two-period changeover design with no carryover and with fixed subject effects. Other typical examples of hierarchical normal linear models in biomedical applications are Hein *et al.* (2005), Lewin *et al.* (2006) for gene expression data, and Müller *et al.* (1999) for case-control studies.

3.2 Dynamic Linear Modeling

In contrast to cross-sectional data, we frequently encounter situations where the responses and covariates are observed sequentially over time. It is of interest to develop inference for such problems in the context of a dynamic linear model. Let $\mathbf{y}_1, \ldots, \mathbf{y}_T$ denote p-dimensional random variables which are available at times $1, \ldots, T$. Suppose $\mathbf{y}_t$ depends on an unknown q-dimensional *state* vector $\boldsymbol{\theta}_t$ (which may again be scalar or vector-valued) via the *observation equation*

$$\mathbf{y}_t = \mathbf{F}_t\boldsymbol{\theta}_t + \mathbf{v}_t \tag{17}$$

where $\mathbf{F}_t$ is a known $p \times q$ matrix, and we assume that the observation error $\mathbf{v}_t \sim N(\mathbf{0}, \mathbf{V}_t)$, with known $\mathbf{V}_t$. The dynamic change in $\boldsymbol{\theta}_t$ is represented by the *state equation*

$$\boldsymbol{\theta}_t = \mathbf{G}_t\boldsymbol{\theta}_{t-1} + \mathbf{w}_t \tag{18}$$

where $\mathbf{G}_t$ is a known $q \times q$ state transition matrix, and the state error $\mathbf{w}_t \sim N(\mathbf{0}, \mathbf{W}_t)$, with known $\mathbf{W}_t$. In addition, we suppose that $\mathbf{v}_t$ and $\mathbf{w}_t$ are independently distributed. Note that θ_t is a random vector; let

$$\boldsymbol{\theta}_t|\mathbf{y}_t \sim N(\widehat{\boldsymbol{\theta}}_t, \boldsymbol{\Sigma}_t) \tag{19}$$

represent the posterior distribution of $\boldsymbol{\theta}_t$. The *Kalman filter* is a recursive procedure for determining the posterior distribution of $\boldsymbol{\theta}_t$ in this conjugate setup, and thereby predicting $\mathbf{y}_t$ (West and Harrison 1989). Estimation of parameters via the EM algorithm (Dempster *et al.* 1977) has been discussed in Shumway and Stoffer (2004), who provide R code for handling such models; see also http://cran.r-project.org/doc/packages/dlm.pdf (R package from G. Petris). Gamerman and Migon (1993) extended this to dynamic hierarchical modeling in the Gaussian framework.

Gordon and Smith (1990) used a dynamic linear model framework to model and monitor medical time series such as body-weight adjusted reciprocal serum creatanine concentrations for online monitoring under renal transplants, and daily white blood count cells levels in patients with chronic kidney disorders. Kristiansen, Sjöström and Nygaard (2005) used the Kalman filter based on a double integrator for tracking urinary bladder filling from intermittent bladder volume measurements taken by an ultrasonic bladder volume monitor. Wu *et al.* (2003) have used a switching Kalman filter model as a real time decoding algorithm for a neural prosthesis application, specifically for the real-time inference of hand kinematics from a population of motor cortical neurons.

3.3 Beyond Linear Modeling

The linear mixed model (LMM) generalizes the fixed effects models discussed above to include random effects and has wide use in biomedical applications. Let $\boldsymbol{\beta}$ and $\mathbf{u}$ denote p-dimensional and q-dimensional location vectors related to the n-dimensional observation vector $\mathbf{y}$ through the regressor matrix $\mathbf{X}$ and design matrix $\mathbf{Z}$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}; \quad \boldsymbol{\varepsilon}|\sigma_\varepsilon^2 \sim N_n(\mathbf{0}, \sigma_\varepsilon^2\mathbf{I}) \tag{20}$$

so that $\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \sigma_\varepsilon^2 \sim N_n(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}, \sigma_\varepsilon^2\mathbf{I})$. Assume priors $\pi(\boldsymbol{\beta}|\sigma_\beta^2) \sim N(\mathbf{0}, \sigma_\beta^2\mathbf{B})$, $\mathbf{u}|\mathbf{V}\sigma_u^2 \sim N(\mathbf{0}, \sigma_u^2\mathbf{V}\sigma_u^2)$, where $\mathbf{B}$ and $\mathbf{V}$ are known, nonsingular matrices, and σ_u^2 and σ_β^2 are unknown hyperparameters. It is easily seen that no closed form expressions for the posterior distributions are possible, and inference for these models is feasible via the sampling based Bayesian approach described in Section 4.

Bayesian analysis for generalized linear models (GLIM's) useful for analyzing non-normal data has seen rapid growth in the last two decades; see Gelfand and Ghosh (2000) for an overview and summary. West (1985) and Albert (1988) were among early discussants of a general hierarchical framework for GLIM's. Most familiar examples are logit or probit models for binary/binomial responses, loglinear models for responses of counts, cumulative logit models for ordinal categorical responses, etc. Suppose the responses $y_1, \ldots, y_n$ are independent with pdf belonging to an exponential family, so that the likelihood is

$$l(\boldsymbol{\theta}|\mathbf{y}) = \prod_{i=1}^{n} \exp[a^{-1}(\phi_i)\{y_i\theta_i - \psi(\theta_i)\} + c(y_i;\phi_i)] \tag{21}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$, θ_i are unknown parameters related to the predictors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and regression parameters $\boldsymbol{\beta}$ via $\theta_i = h(\mathbf{x}_i'\boldsymbol{\beta})$ for a strictly increasing sufficiently smooth *link* function $h(.)$, and $a(\phi_i)$ is known. A simple conjugate prior for $\boldsymbol{\beta}$ is $N(\boldsymbol{\beta}_0, \boldsymbol{\Sigma})$ where $\boldsymbol{\beta}_0$ and $\boldsymbol{\Sigma}$ are known, so that the posterior has the form

$$\pi(\boldsymbol{\beta}|\mathbf{y}) \propto \exp\{\sum_i a^{-1}(\phi_i)[y_i h(\mathbf{x}_i'\boldsymbol{\beta}) - \psi(h(\mathbf{x}_i'\boldsymbol{\beta}))] - \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\} \tag{22}$$

This posterior is not analytically tractable. Computational Bayesian methods discussed in Section 4 enable inference in complex situations such as generalized linear mixed models (GLMM's) (Clayton 1996), nonlinear random effects models (Dey, Chen and Chang 1997), models with correlated and strongly correlated random effects and hierarchical GLMM's (Sun, Speckman and Tsutakawa, 2000), correlated categorical responses (Chen and Dey 2000), overdispersed GLIM's (Dey and Ravishanker 2000), and survival data models (Kuo and Peng 2000). Applications of such models to disease maps is discussed further in Section 6 (see also Waller *et al.* 1997).

4 Computational Bayesian Framework

As indicated in the previous discussion, the main ingredients of the Bayesian feast are probabilistic models (parametric or semi-parametric) and priors distributions, which when combined, produce posterior distributions, predictive distributions and summaries thereof. More specifically, recall the posterior, the predictive and the posterior predictive distributions (equations (4), (9) and (8)), i.e., $\pi(\boldsymbol{\theta}|\mathbf{y}) = \pi(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})/m(\mathbf{y})$, $m(\mathbf{y}) = \int \pi(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta}$ and $p(\widetilde{y}|\mathbf{y}) = \int p(\widetilde{y}|\boldsymbol{\theta},\mathbf{y})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$. The Bayesian agenda includes, among other things, posterior modes, $\max_{\boldsymbol{\theta}} \pi(\boldsymbol{\theta}|\mathbf{y})$, posterior moments, $E_\pi[g(\boldsymbol{\theta})]$, density estimation, $\hat{\pi}(g(\boldsymbol{\theta})|\mathbf{y})$, Bayes factors, $m_0(\mathbf{y})/m_1(\mathbf{y})$, and decision making, $\max_d \int U(d,\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$.

Historically, those tasks were (partially) performed by analytic approximations, which include asymptotic approximations (Carlin and Louis 2000), Gaussian quadrature (Naylor and Smith 1982) and Laplace approximations (Tierney and Kadane 1986; and Kass, Tierney and Kadane 1988; Tierney, Kass and Kadane 1989). Modern, fast and cheap computational resources have facilitated the widespread use of Monte Carlo methods, which in turn has made Bayesian reasoning commonplace in almost every area of scientific research. This trend is overwhelmingly apparent in the biomedical sciences (Do, Müller and Vannucci 2006; Larget 2005; Sorensen and Gianola 2002). Initially, simple Monte Carlo schemes were extensively used for solving practical Bayesian problems, including the Monte Carlo method (Geweke 1989), the rejection algorithm (Gilks and Wild 1992), the weighted resampling algorithm (Smith and Gelfand 1992), among others. Currently, one could argue that the most widely used Monte Carlo schemes are the Gibbs sampler/data augmentation algorithm (Gelfand and Smith 1990; Tanner and Wong 1987) and the Metropolis-Hastings algorithm (Metropolis *et al.* 1953; Hastings 1970), which fall in the category of Markov Chain Monte Carlo algorithms (Gilks, Richardson and Spiegelhalter 1996). This section only briefly introduces the main algorithms. The more curious and energetic reader will find in Gamerman and Lopes (2006) and its associated web-page `www.ufrj.br/mcmc`, among other things, extensions of these basic algorithms, recent developments, didactic examples and their R codes and an extensive and updated list of freely downloadable statistical routines and packages.

4.1 Normal Approximation

Let $\mathbf{m}$ be the posterior mode, i.e., $\mathbf{m} = \arg\max_{\boldsymbol{\theta}} \pi(\boldsymbol{\theta}|\mathbf{y})$; a standard Taylor series expansion leads to approximating $\pi(\boldsymbol{\theta}|\mathbf{y})$ by a (multivariate) normal distribution with mean vector $\mathbf{m}$ and precision matrix $\mathbf{V}^{-1} = -\frac{\partial^2 \log \pi(\mathbf{m}|\mathbf{y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}$. This approximation can be thought of as a Bayesian version of the central limit theorem (Carlin and Louis 2000; Heyde and Johnstone 1979; Schervish 1995). Finding $\mathbf{m}$ is not a trivial task and generally involves solving a set of nonlinear equations, usually by means of iterative Newton-Raphson-type and Fisher's scoring algorithms (Thisted 1988). For most problems, specially in high dimension, normal approximations tend to produce rather crude and rough estimates.

4.2 Integral Approximation

One of the main tasks in Bayesian analysis is the computation of posterior summaries, such as means, variances and other moments, of functions of $\boldsymbol{\theta}$, say $t(\boldsymbol{\theta})$, i.e.,

$$E[t(\boldsymbol{\theta})] = \frac{\int t(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int p(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} \tag{23}$$

In this section, a brief review of the main analytic and stochastic approximations to the above integral is provided.

Quadrature approximation

Quadrature rules approximate unidimensional integrals $\int_a^b g(\theta)d\theta$, for instance $g(\theta) = t(\theta)p(y|\theta)\pi(\theta)$ in equation (23), by $\sum_{i=1}^n w_i g(\theta_i)$ for suitably chosen weights w_i and grid points θ_i, $i = 1, \ldots, n$. Simple quadrature rules are Simpson's and the trapezium rules. *Gaussian quadrature* are special rules for situations where $g(\theta)$ can be well approximated by the product of a polynomial and a density function. Gauss-Jacobi rule arises under the uniformity $[-1, 1]$ density, while Gauss-Laguerre and Gauss-Hermite rules arise under gamma and normal densities. See, for instance, Abramowitz and Stegun (1965) for tabulations and Naylor and Smith (1982) and Pole and West (1990) for Bayesian inference under quadrature. Quadrature rules are not practical even in problems of moderate dimensions.

Laplace approximation

For $t(\boldsymbol{\theta}) > 0$, Equation (23) can be written in the exponential form as

$$E[t(\boldsymbol{\theta})] = \frac{\int \exp\{L^*(\boldsymbol{\theta})\}d\boldsymbol{\theta}}{\int \exp\{L(\boldsymbol{\theta})\}d\boldsymbol{\theta}} \tag{24}$$

where $L^*(\boldsymbol{\theta}) = \log t(\boldsymbol{\theta}) + \log p(\mathbf{y}|\boldsymbol{\theta}) + \log \pi(\boldsymbol{\theta})$ and $L(\boldsymbol{\theta}) = \log p(\mathbf{y}|\boldsymbol{\theta}) + \log \pi(\boldsymbol{\theta})$. Better (than normal) approximations can be obtained by Taylor series expansions both in the numerator and the denominator of the previous equation. It can be shown under fairly general conditions, and when $t(\boldsymbol{\theta}) > 0$ that

$$\hat{E}_{lap}[t(\boldsymbol{\theta})] = \left(\frac{|\mathbf{V}^*|}{|\mathbf{V}|}\right)^{1/2} \exp\left\{L^*(\mathbf{m}^*) - L(\mathbf{m})\right\}. \tag{25}$$

where $\mathbf{m}^*$ is the value of $\boldsymbol{\theta}$ that maximizes L^* and $\mathbf{V}^*$ as minus the inverse Hessian of L^* at the point $\mathbf{m}^*$. This approximation is known as the *Laplace approximation* (Kass, Tierney and Kadane 1988). Laplace approximations for the case where $t(\boldsymbol{\theta}) < 0$ appear in Tierney, Kass and Kadane (1989). The Laplace approximation tends to be poor when either the posterior is multimodal or when approximate normality fails.

Monte Carlo integration

If a sample $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ from the prior $\pi(\boldsymbol{\theta})$ is available, then

$$\hat{E}_{mc}[t(\boldsymbol{\theta})] = \frac{\sum_{j=1}^n t(\boldsymbol{\theta}_j)p(\mathbf{y}|\boldsymbol{\theta}_j)}{\sum_{j=1}^n p(\mathbf{y}|\boldsymbol{\theta}_j)} \tag{26}$$

is a *Monte Carlo* (MC) estimator of $E[t(\boldsymbol{\theta})]$ in Equation (23). Limiting theory assures us that, under mild conditions on $t(\boldsymbol{\theta})$, the above MC estimator converges to its mean $E[t(\boldsymbol{\theta})]$ (Geweke 1989).

It is well known that sampling from the prior distribution may produce poor estimates and a practically infeasible number of draws will be necessary to achieve reasonably accurate levels of approximation. The MC estimator (26) can be generalized for situations where, roughly speaking, a proposal density $q(\boldsymbol{\theta})$ is available that mimics $\pi(\boldsymbol{\theta}|\mathbf{y})$ in the center, dominates it in the tails, and is easy to sample from. More specifically, if a sample $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ from $q(\boldsymbol{\theta})$ is available (see sampling from distributions later in this section), then another MC estimator for Equation (23) is

$$\hat{E}_{mcis}[t(\boldsymbol{\theta})] = \frac{\sum_{j=1}^n t(\boldsymbol{\theta}_j)p(\mathbf{y}|\boldsymbol{\theta}_j)\pi(\boldsymbol{\theta}_j)/q(\boldsymbol{\theta}_j)}{\sum_{j=1}^n p(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}_j)/q(\boldsymbol{\theta}_j)}. \tag{27}$$

This estimator is commonly known as a *Monte Carlo via Importance Sampling* (MCIS) estimator of $E[t(\boldsymbol{\theta})]$. It is easy to see that the previous MC estimator is a special case of the MCIS when $q(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})$. The proposal density q is also referred to as *importance density* and sampling from q is known as *importance sampling*. Limiting theory assures us that MC estimators, under mild conditions on $t(\boldsymbol{\theta})$ and $q(\boldsymbol{\theta})$, approximate the actual posterior expectations (see Geweke 1989)

4.3 Monte Carlo-based Inference

Based on the previous argument, if $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ is a readily available sample from the posterior $\pi(\boldsymbol{\theta}|\mathbf{y})$ then the MC approximation to $E[t(\boldsymbol{\theta})]$ would be $n^{-1}\sum_{j=1}^{n} t(\boldsymbol{\theta}_j)$. The problem is that $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ is rarely readily available. We now discuss two algorithms for sampling independent draws from $\pi(\boldsymbol{\theta}|\mathbf{y})$, viz., the *rejection algorithm* and the *weighted resampling algorithm*. We also discuss two iterative algorithms, viz., the *Gibbs sampler* and the *Metropolis-Hastings algorithm*, that sample from Markov chains whose limiting, equilibrium distributions are the posterior $\pi(\boldsymbol{\theta}|\mathbf{y})$. One common aspect of all these algorithms is that they all potentially use draws from auxiliary, proposal, importance densities, $q(\boldsymbol{\theta})$, whose importance are weighted against the target, posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y})$, i.e., by considering weights $\pi(\boldsymbol{\theta}|\mathbf{y})/q(\boldsymbol{\theta})$ (see the explanation between Equations (26) and (27)). For notational reasons, let $\tilde{\pi}(\boldsymbol{\theta}) = p(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ be the unnormalized posterior, while $\omega(\boldsymbol{\theta}) = \tilde{\pi}(\boldsymbol{\theta})/q(\boldsymbol{\theta})$ is the unnormalized weight.

Rejection method

When samples are easily drawn from a proposal q such that $\tilde{\pi}(\boldsymbol{\theta}) \leq Aq(\boldsymbol{\theta})$, for some finite A and for all possible values of $\boldsymbol{\theta}$, then it can be shown that the following algorithm produces independent draws from $\tilde{\pi}(\boldsymbol{\theta})$. The proposal density q is commonly known as a *blanketing density* or an *envelope density*, while A is the *envelope constant*. If both $\tilde{\pi}$ and q are normalized densities, then $A \geq 1$ and the theoretical acceptance rate is given by $1/A$. In other words, more draws are likely to be accepted the closer q is to π, i.e., the closer A is to one.

Algorithm 1. Rejection method

1. Set $j = 1$;
2. Draw $\boldsymbol{\theta}^*$ from q and u from $U[0, 1]$;
3. Compute the unnormalized weight, $\omega(\boldsymbol{\theta}^*)$;
4. If $Au \leq \omega(\boldsymbol{\theta}^*)$, then set $\boldsymbol{\theta}_j = \boldsymbol{\theta}^*$ and $j = j + 1$;
5. Repeat Steps 2, 3 and 4 while $j \leq n$.

The resulting sample $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n\}$ is distributed according to $\pi(\boldsymbol{\theta}|\mathbf{y})$.

Weighted resampling method

When A is not available or is hard to derive, *weighted resampling* is a direct alternative since it can use draws from a density q without having to find the constant A. This algorithm is more commonly known as the *sampling importance resampling* (SIR) algorithm (Smith and Gelfand 1992). We should also mention other adaptive rejection algorithms and the renewed interest in SIR-type algorithms in sequential Monte Carlo (Doucet, de Freitas and Gordon 2001) and population Monte Carlo (Cappé *et al.* 2004).

Algorithm 2. Weighted resampling method

1. Sample $\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_m^*$ from q;
2. For $i = 1, \ldots, m$
   (a) Compute unnormalized weights: $\omega_i = \omega(\boldsymbol{\theta}_i^*)$;
   (b) Normalize weights: $w_i = \frac{\omega_i}{\sum_{l=1}^{m} \omega_i}$;
3. Sample $\boldsymbol{\theta}_j$ from $\{\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_m^*\}$, such that $Pr(\boldsymbol{\theta}_j = \boldsymbol{\theta}_i^*) = w_i$, for $j = 1, \ldots, n$.

For large m and n, the resulting sample $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n\}$ is approximately distributed according to $\pi(\boldsymbol{\theta}|\mathbf{y})$.

Metropolis-Hastings algorithm

Instead of discarding the current rejected draw, as in the rejection algorithm, Metropolis-Hastings (MH) algorithms (Metropolis *et. al.* 1953; Hastings 1970) translate the rejection information into higher importance, or weight, to the previous draw. This generates an iterative chain of dependent draws (a Markov scheme). Markov chain arguments guarantee that, under fairly general regularity conditions and in the limit, such a Markov scheme generates draws from π, also known as the *target, limiting, equilibrium* distribution of the chain (Tierney 1994).

Algorithm 3. Metropolis-Hastings algorithm

1. Set the initial value at $\boldsymbol{\theta}_0$ and $j = 1$;
2. Draw $\boldsymbol{\theta}^*$ from $q(\boldsymbol{\theta}_{j-1}, \cdot)$ and u from $U[0, 1]$;
3. Compute unnormalized weights and the acceptance probability

$$\begin{aligned} \omega(\boldsymbol{\theta}_{j-1}, \boldsymbol{\theta}^*) &= \tilde{\pi}(\boldsymbol{\theta}^*)/q(\boldsymbol{\theta}_{j-1}, \boldsymbol{\theta}^*) \\ \omega(\boldsymbol{\theta}^*, \boldsymbol{\theta}_{j-1}) &= \tilde{\pi}(\boldsymbol{\theta}_{j-1})/q(\boldsymbol{\theta}^*, \boldsymbol{\theta}_{j-1}) \\ \alpha(\boldsymbol{\theta}_{j-1}, \boldsymbol{\theta}^*) &= \min\left\{1, \frac{\omega(\boldsymbol{\theta}_{j-1}, \boldsymbol{\theta}^*)}{\omega(\boldsymbol{\theta}^*, \boldsymbol{\theta}_{j-1})}\right\}; \end{aligned}$$

4. If $u \leq \alpha(\boldsymbol{\theta}_{j-1}, \boldsymbol{\theta}^*)$ set $\boldsymbol{\theta}_j = \boldsymbol{\theta}^*$, otherwise set $\boldsymbol{\theta}_j = \boldsymbol{\theta}_{j-1}$;
5. Set $j = j + 1$ and go back to Step 2 until convergence is reached.

In the above algorithmic representation, the proposal density $q(\cdot, \boldsymbol{\theta})$ plays a similar role as $q(\boldsymbol{\theta})$ in both rejection and SIR algorithms. Two commonly used versions of the MH algorithm are the *random walk* MH and the *independent* MH, where $q(\boldsymbol{\phi}, \boldsymbol{\theta}) = q(|\boldsymbol{\theta} - \boldsymbol{\phi}|)$ and $q(\boldsymbol{\phi}, \boldsymbol{\theta}) = q(\boldsymbol{\theta})$, respectively. Proper choice of q and convergence diagnostics are of key importance to validate the algorithm and are, in the majority of the situations, problem-specific. See Chen Shao and Ibrahim (2000) or Gamerman and Lopes (2006), Chapter 6, for additional technical details, references and didactic examples.

Gibbs sampler

Like MH algorithms, the *Gibbs sampler* is an iterative MC algorithm that takes advantage of easy to sample from, and easy to evaluate full conditional distributions that appear in several statistical modeling structures. More specifically, it is an algorithm that breaks the vector of

parameters $\boldsymbol{\theta}$ into d blocks (scalar, vector or matrix) of parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d$ and recursively samples $\boldsymbol{\theta}_i$ from its full conditional $\pi(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{-i}, \mathbf{y})$, where $\boldsymbol{\theta}_{-i} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{i-1}, \boldsymbol{\theta}_{i+1}, \ldots, \boldsymbol{\theta}_d)$, for $i = 1, \ldots, d$. Its name is derived from its initial use in the context of image processing, where the posterior distribution was a Gibbs distribution (Geman and Geman 1984), while Gelfand and Smith (1990) made it popular within the statistical community. As the number of $\boldsymbol{\theta}$ draws increases, the chain approaches its equilibrium, viz., $\pi(\boldsymbol{\theta}|\mathbf{y})$. Convergence is then assumed to hold approximately. It can be described algorithmically as follows.

Algorithm 4. The Gibbs sampler

1. Set the initial values at $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\theta}_1^{(0)}, \ldots, \boldsymbol{\theta}_d^{(0)})'$ and $j = 1$;
2. Obtain a new value $\boldsymbol{\theta}^{(j)} = (\boldsymbol{\theta}_1^{(j)}, \ldots, \boldsymbol{\theta}_d^{(j)})'$ from $\boldsymbol{\theta}^{(j-1)}$ as follows

$$\begin{aligned}
\boldsymbol{\theta}_1^{(j)} &\sim \pi(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2^{(j-1)}, \ldots, \boldsymbol{\theta}_d^{(j-1)}) , \\
\boldsymbol{\theta}_2^{(j)} &\sim \pi(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1^{(j)}, \boldsymbol{\theta}_3^{(j-1)}, \ldots, \boldsymbol{\theta}_d^{(j-1)}) , \\
&\vdots \\
\boldsymbol{\theta}_d^{(j)} &\sim \pi(\boldsymbol{\theta}_d|\boldsymbol{\theta}_1^{(j)}, \ldots, \boldsymbol{\theta}_{d-1}^{(j)}) ;
\end{aligned}$$

3. Set $j = j + 1$ and go back to 2 until convergence is reached.

MCMC over model spaces

Suppose models M_j, for $j \in \mathcal{J}$ (for instance, $\mathcal{J} = \{1, \ldots, J\}$), are entertained. Also, under model M_j, assume that $p(\mathbf{y}|\theta_j, M_j)$ is a probability model for $\mathbf{y}$ parameterized by a d_j-dimensional vector $\boldsymbol{\theta}_j$ (usually in $\Re^{d_j}$). Because the dimension (and interpretation) of the model parameters might vary with the models, the above MCMC algorithms cannot be directly used in order to derive, for example, approximations for $\pi(\boldsymbol{\theta}_j|\mathbf{y}, M_j)$ and, perhaps more importantly, $Pr(M_j|\mathbf{y})$, the posterior model probability for model M_j. Nonetheless, Carlin and Chib (1995) and Green (1995) respectively proposed generalizations of the Gibbs sampler and the Metropolis-Hastings algorithm to situations where the parameter vector becomes $(j, \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\boldsymbol{\theta}_j : j \in \mathcal{J})$. The former is commonly known as the *Carlin-Chib algorithm* and the latter as the *reversible jump* MCMC (RJMCMC) algorithm. Dellaportas, Forster and Ntzoufras (2002) and Godsill (2001) propose hybrid versions of these two algorithms, while Clyde (1999) argues that MCMC model composition (Raftery, Madigan and Hoeting 1997) and stochastic search variable selection (George and McCulloch, 1992) are particular cases of the reversible jump MCMC (RJMCMC) algorithm. Amongst many others, Kuo and Song (2005) used RJMCMC for carrying out inference in dynamic frailty models for multivariate survival times (see Section 5.3), Waagepetersen and Sorensen (2001) used it in genetic mapping and Lopes, Müller and Rosner (2003) in multivariate mixture modeling of hematologic data (see Example 2). Detailed and comprehensive review of MCMC algorithms over model spaces appeared in Sisson (2005) along with an extensive list of freely available packages for RJMCMC and other transdimensional algorithms. Further details appear in Chapter 7 of Gamerman and Lopes (2006).

Public domain software

Until recently, a practical impediment to the routine use of Bayesian approaches was the lack of reliable software. This has rapidly changed over the last few years. The BUGS project (Spiegelhalter, Thomas and Best 1999) provides public domain code that has greatly facilitated the use of

Bayesian inference. The BUGS code is available from `http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/conten`
The software is widely used and well tested and validated.

Another source of public domain software for Bayesian inference consists of libraries for the statistical analysis system **R** (R Development Care Team 2006). **R** provides many libraries (packages) for Bayesian inference, including *MCMCpack* for MCMC within R, *bayesSurv* for Bayesian survival regression, *boa* for MCMC convergence diagnostics, *DPpackage* for non-parametric Bayesian inference, and *BayesTree* for Bayesian additive regression trees. All packages are accessible and can be downloaded from the main **R** website `http://cran.r-project.org`.

5 Bayesian Survival Analysis

Bayesian methods for survival analysis to model times-to-events have been widely used in recent years in the biomedical and public health areas. Ibrahim, Chen and Sinha (2001) is an excellent text describing various aspects of Bayesian inference. Here, we give a brief summary, with references, of a few useful areas, and then describe in detail frailty models for multivariate survival times.

5.1 Models for Univariate Survival Times

Univariate survival analysis assumes a suitable model for a continuous nonnegative survival time T which may be described in terms of the survival function $S(t) = P(T > t)$ or the hazard function $h(t) = -d \log S(t)/dt$. A parametric framework assumes that i.i.d. survival times $\mathbf{t} = (t_1, \ldots, t_n)$ follow a parametric model such as exponential, Weibull, gamma, lognormal, or poly-Weibull. The data might be complete or censored. Recall that a survival time is right (left) censored at c if its actual value is unobserved and it is only known that the time is greater than or equal to (less than or equal to) c, and is interval censored if it is only known that it lies in the interval (c_1, c_2). Given a set of p covariates (risk factors) $Z_1, \ldots, Z_p$, it is straightforward to use a standard software like BUGS to fit a suitable regression model and derive posterior and predictive distributions. A proportional hazards model specifies $h(t|\mathbf{z}) = h_0(t) \exp(\mathbf{z}_i' \boldsymbol{\beta})$ for a vector of parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$, and baseline hazard $h_0(.)$. Biomedical applications of parametric models have been discussed for instance in Achcar, Brookmeyer and Hunter (1985) and Kim and Ibrahim (2000). For situations in which the proportional hazards may be inappropriate, flexible hierarchical models with time dependent covariates have been fit so that $h(t|\mathbf{z}) = h_0(t) \exp(\mathbf{x}_i' \boldsymbol{\beta}(t)$, or even by replacing the linear term by a neural network; see Gustafson (1998), Carlin and Hodges (1999) and Faraggi and Simon (1995). Gustafson (1998) discusses data on duration of nursing home stays via the hybrid MCMC algorithm. As another alternative, a generalization of the Cox model has been discussed in Sinh, Chen and Gosh (1999) via a discretize hazard and time dependent regression coefficients, with application to breast cancer data (Finkelstein and Wolfe 1985).

With recent advances in computational Bayesian methods, semiparametric and nonparametric methods have become more prevalent. The piecewise constant semiparametric hazard model defines a finite partition of the time axis into K intervals, and assumes a constant baseline hazard in each subinterval; the simplest baseline, called the piecewise exponential model, will be described under frailty modeling later in this section. Nonparametric models include use of gamma process priors (Kalbfleisch 1978), beta process priors (Sinha 1997), correlated gamma process priors (Mezzetti and Ibrahim 2000), Dirichlet process priors (DPP) (Gelfand and Mallick 1995), mixtures of DPP's or MDP models (MacEachern and Müller 1998), and Polya tree process priors (Lavine 1992).

5.2 Shared Frailty Models for Multivariate Survival Times

Dependent multivariate times to events frequently occur in several biomedical applications, and frailty models (Vaupel, Manton and Stallard 1979) have been extensively used for modeling dependence in such multivariate survival data. The dependence frequently arises because subjects in the same group are related to each other or due to multiple recurrence times of a disease for the same patient, and computational Bayesian methods facilitate a variety of frailty models. The widely used shared frailty models for multivariate times to events assume that there is an unobserved random effect, known as frailty, which explains dependence that may arise due to association among subjects in the same group, or among multiple recurrence times of an event for the same subject. Suppose the survival time of the kth subject $(k = 1, \ldots, m)$ in the jth group $(j = 1, \ldots, n)$ is denoted by T_{jk}, and $\mathbf{z}_{jk}$ is a fixed, possibly time dependent covariate vector of dimension p.

Example 5 *An often cited example (McGilchrist and Aisbett 1991) consists of the kidney infection data on times to first and second occurrence of infection in 38 patients on portable dialysis machines ($n = 38$, $m = 2$). Binary variables representing respectively the censoring indicators for the first and second recurrences are available; occurrence of infection is indicated by 1, and censoring by 0. The gender of the patients (0 indicating male, and 1 indicating female), is a covariate (see Ibrahim, Chen and Sinha 2001, Table 1.5). Other covariates, such as age and disease type of each patient, are also available with this data. However, initial analysis by McGilchrist and Aisbett, using what they referred to as a penalized partial likelihood approach, showed that the effect of these covariates on infection times was not statistically significant; hence they are omitted from the analysis.*

Given the unobserved frailty parameter w_j for the jth group, the modified Cox proportional hazards shared frailty regression model (Clayton and Cuzick 1985; Oakes 1989) via the hazard function is

$$h(t_{jk}|w_j, \mathbf{z}_{jk}) = h_0(t_{jk}) \exp(\boldsymbol{\beta}'\mathbf{z}_{jk}) w_j \tag{28}$$

where $\boldsymbol{\beta}$ is the vector of regression parameters of the same dimension, $h_0(.)$ is the baseline hazard function and w_j is an individual random effect (frailty) representing common unobserved covariates and generating dependence. The event times are assumed to be conditionally independent given the shared frailty. For identifiability purposes, it is usual to require that the linear model component $\boldsymbol{\beta}'\mathbf{z}_{jk}$ has no intercept term. Let $\theta_{jk} = \exp(\boldsymbol{\beta}'\mathbf{z}_{jk})$.

Baseline hazard function

The baseline hazard function could be a simple constant hazard function, a Lévy process, a Gamma process, a Beta process or a correlated prior process (see Sinha and Dey 1997 for an extensive review). A common parametric baseline hazard is the Weibull form $h_0(t_{ij}) = \lambda\gamma t_{ij}^{\gamma-1}$, for $\lambda > 0$, and $\gamma > 0$. The more flexible piecewise exponential correlated prior process baseline hazard requires that the time period is divided into g intervals, $I_i = (t_{i-1}, t_i)$ for $i = 1, \ldots, g$, where $0 = t_0 < t_1 < \ldots < t_g < \infty$, t_g denoting the last survival or censored time. The baseline hazard is assumed to be constant within each interval, i.e., $\lambda_0(t_{jk}) = \lambda_i$ for $t_{jk} \in I_i$. A discrete-time martingale process is used to correlate the λ_i's in adjacent intervals, thus introducing some smoothness (Arjas and Gasbarra 1994). Given $(\lambda_1, \ldots, \lambda_{i-1})$, specify that

$$\lambda_i|\lambda_1, \ldots, \lambda_{i-1} \sim \text{Gamma}\left(c_i, \tfrac{c_i}{\lambda_{i-1}}\right), i = 1, \ldots, g$$

where $\lambda_0 = 1$, so that $E(\lambda_i|\lambda_1, \ldots, \lambda_{i-1}) = \lambda_{i-1}$; let $\tilde{\lambda} = (\lambda_1, \ldots, \lambda_g)$. A small value of c_i indicates less information for smoothing the λ_i's; if $c_i = 0$, then λ_i is independent of λ_{i-1} while if $c_i \to \infty$, $\lambda_i = \lambda_{i-1}$. Regarding the choice of g, a very large value would result in a non-parametric model

and produce unstable estimates of λ's while a very small value of g would lead to inadequate model fitting. In practical situations, g is determined based on the design. A random choice of g will lead to a posterior distribution with variable dimensions, which may be handled via a reversible jump MCMC (Green, 1995) described in Section 4.

Frailty specifications

Alternate parametric shared frailty specifications that have appeared in the recent literature include the gamma model (Clayton and Cuzick 1985), the log-normal model (Gustafson 1997), and the positive stable model (Hougaard 2000). The gamma distribution is the most common finite mean frailty distribution, and we assume that w_j are i.i.d. Gamma$(\kappa^{-1}, \kappa^{-1})$ variables, so that the mean is 1 and the variance is the unknown κ. For identifiability, the mean of w_j's must be 1. For Bayesian inference on a shared Gamma frailty-Weibull baseline model with application to the kidney infection data, see Section 4.1 in Ibrahim *et al.* (2001). In a proportional hazards frailty model, the unconditional effect of a covariate, which is measured by the hazard ratio between unrelated subjects (i.e., with different frailties) is always less than its conditional effect, measured by the hazard ratio among subjects with the same frailty. In particular, suppose we consider two subjects from different groups and with respective covariates 0 and $\tilde{z}$; let $S_0(t)$ and $S_1(t)$ denote the corresponding unconditional survivor functions derived under this frailty specification. It has been shown that the covariate effects, as measured by the hazard ratio are always attenuated and further, $S_0(t)$ and $S_1(t)$ are usually not related via a proportional hazards model. If the frailty distribution is an infinite variance positive stable distribution, then $S_1(t)$ and $S_0(t)$ will have proportional hazards; we no longer need to choose between conditional and unconditional model specifications, since a single specification can be interpreted either way. A positive stable frailty distribution thus not only permits a proportional hazards model to apply unconditionally, but also allows for a much higher degree of heterogeneity among the common covariates than would be possible by using a frailty distribution with finite variance, such as the Gamma distribution. The frailty parameters w_j, $j = 1, \ldots, n$ are assumed to be independent and identically distributed for every group, according to a positive α- stable distribution. The density function of a positive stable random variable w_j is not available in closed form. However, its characteristic function is available and has the form

$$E(e^{i\vartheta w_j}) = \exp\{-|\vartheta|^{\alpha}(1 - i\text{sign}(\vartheta)\tan(\pi\alpha/2))\} \tag{29}$$

where $i = \sqrt{-1}$, ϑ is a real number, sign$(\vartheta) = 1$ if $\vartheta > 0$, sign$(\vartheta) = 0$ if $\vartheta = 0$ and sign$(\vartheta) = -1$ if $\vartheta < 0$.

Positive stable shared frailty model

Although the positive stable frailty model is conceptually simple, estimation of the resulting model parameters is complicated due to the lack of a closed form expression for the density function of a stable random variable. Qiou, Ravishanker and Dey (1999) described a Bayesian framework using MCMC for this problem with application to the kidney infection data. This was later extended by Ravishanker and Dey (2000) to include a mixture of positive stables frailty, and by Mallick and Ravishanker (2004) to a power variance family (PVF) frailty (indexed by parameters η and α). The Bayesian approach is based on an expression provided by Buckle (1995) for the joint density of n i.i.d. observations from a stable distribution by utilizing a bivariate density function $f(w_j, y|\alpha)$ whose marginal density with respect to one of the two variables gives exactly a stable density. Let $f(w_j, y|\alpha)$ be a bivariate function such that it projects $[(-\infty, 0) \times (-1/2, l_\alpha)] \cup [(0, \infty) \times (l_\alpha, 1/2)]$ to $(0, \infty)$:

$$f(w_j, y|\alpha) = \frac{\alpha}{|\alpha - 1|} \exp\{-|\frac{w_j}{\tau_\alpha(y)}|^{\alpha/(\alpha-1)}\}|\frac{w_j}{\tau_\alpha(y)}|^{\alpha/(\alpha-1)} \frac{1}{|w_j|} \tag{30}$$

where

$$\tau_\alpha(y) = \frac{\sin(\pi\alpha y + \psi_\alpha)}{\cos \pi y}[\frac{\cos \pi y}{\cos\{\pi(\alpha-1)y + \psi_\alpha\}}]^{(\alpha-1)/\alpha},$$

$w_j \in (-\infty, \infty)$, $y \in (-1/2, 1/2)$, $\psi_\alpha = \min(\alpha, 2-\alpha)\pi/2$ and $l_\alpha = -\psi_\alpha/\pi\alpha$. Then

$$f(w_j|\alpha) = \frac{\alpha|w_j|^{1/(\alpha-1)}}{|\alpha-1|}\int_{-1/2}^{1/2} \exp\{-|\frac{w_j}{\tau_\alpha(y)}|^{\alpha/(\alpha-1)}\}|\frac{1}{\tau_\alpha(y)}|^{\alpha/(\alpha-1)}dy. \quad (31)$$

Denoting by $\mathcal{D}$ the triplets, $(t_{jk}, \delta_{jk}, \mathbf{z}_{jk})$, the vector of unobserved w_j's by $\mathbf{w}$ and the vector of unobserved auxiliary variables $(y_1, \ldots, y_n)$ by $\mathbf{y}$, the complete data likelihood is

$$\begin{aligned} l(\boldsymbol{\beta}, \boldsymbol{\lambda}, \alpha|\mathbf{w}, \mathbf{y}, \mathcal{D}) &= \prod_{j=1}^{n}\prod_{k=1}^{m}[\prod_{i=1}^{g-1}\exp\{-\lambda_i\Delta_i\theta_{jk}w_j\}] \\ &\times \exp\{-\lambda_g(t_{jk}-t_{g-1})\theta_{jk}w_j\}(\lambda_g\theta_{jk}w_j)^{\delta_{jk}}. \end{aligned} \quad (32)$$

The observed data likelihood based on the observed data $\mathcal{D}$ is obtained by integrating out the w_j's from (32) using the stable density expression in (31), and corresponds to the marginal model whereas (5) corresponds to the conditional model given the frailty. Assuming suitable priors, the joint posterior density based on the observed data likelihood is derived and appropriate MCMC algorithms (see Section 4) are used to generate samples from the posterior distribution via complete conditional distributions. For instance, the ratio-of-uniforms algorithm is used to generate λ_k and β_j samples, the Metropolis-Hastings algorithm is used for α, and the rejection algorithm for y_i. See (Qiou, Ravishanker and Dey 1999) for details, as well as for results corresponding to bivariate times with the kidney infection data. In Table 1, we show results from fitting the different shared frailty models to the kidney infection data.

Table 1. Posterior Summary for Kidney Infection Data under Shared Frailty Models

	Frailty Model			
	Gamma	Positive Stable	PVF	Additive Stable
Parameter	mean(s.d.)	mean(s.d.)	mean(s.d.)	mean(s.d.)
β	-1.62(.42)	-1.06(.36)	-1.15(.88)	-1.40(.61)
$1/\kappa$	.33(.17)	-	-	-
α	-	.86(.07)	.38(.35)	-
η	-	-	1.09(1.05)	-
$\alpha 1$	-	-	-	.35(.05)
$\alpha 2$	-	-	-	.39(.04)
$\alpha 3$	-	-	-	.42(.04)
λ_1	.002(.003)	0.001(.002)	.001(.004)	.012(.015)
λ_2	.001(.003)	0.003(.007)	.004(.016)	.023(.036)
λ_3	.001(.002)	0.0038	.005(.019)	.031(.049)
λ_4	.001(.002)	0.004(.009)	.005(.019)	.037(.057)
λ_5	.001(.002)	0.003(.008)	.005(.019)	.040(.060)
λ_6	.001(.002)	0.004(.011)	.005(.019)	.045(.066)
λ_7	.001(.002)	0.004(.010)	.006(.020)	.058(.065)
λ_8	.001(.003)	0.006(.013)	.008(.025)	.097(.098)
λ_9	.006(.005)	0.027(.022)	.038(.063)	.298(.196)
λ_{10}	.360(.150)	1.84(.604)	2.60(2.13)	2.86(1.22)

5.3 Extensions of Frailty Models

There are several extensions of frailty modeling. Unlike the shared frailty model which corresponds to the assumption of a common risk dependence among multivariate times, the additive frailty model (Hougaard 2000) allows us to handle such survival times with varying degrees of dependence, by combination of subgroups. For instance, data on time to tumorigenesis of female rats in litters was discussed by Mantel, Bohidar and Ciminera (1977). Each litter had three rats; one rat was drug-treated and the other two served as control. Time to tumor appearance was recorded (death due to other causes was considered as censoring), and the study was ended after 104 weeks. Under an additive frailty model, it is assumed that the three female rats in each of 50 litters correspond to three different frailty components instead of sharing a single random component under the shared frailty model, thereby yielding a richer dependence structure. Each subcomponent of the resulting multivariate frailty random variable is further decomposed into independent additive frailty variables, and the frailty component of each rat in every litter is the sum of the litter effect and the individual rat effect. The dependence among rats in each litter then arises due to the litter effect, while the individual rat effect generates additional variability. Specifically, the additive frailty model specifies that the components of a frailty random vector are combined additively for the jth subject within the ith group, and they then act multiplicatively in the Cox proportional hazards model, *i.e.*,

$$h(t_{ij}|w_{ij}, \mathbf{z}_{ij}) = \lambda_0(t_{ij}) \exp(\beta' \mathbf{z}_{ij}) w_{ij}; \tag{33}$$

the dependence is generated by setting

$$w_{ij} = \mathbf{A}'_{ij} \mathbf{X}_i, \tag{34}$$

where $\mathbf{X}'_i = (X_{i1}, X_{i2}, \ldots, X_{is})$, and $\mathbf{A}'_{ij} = (a_{ij1}, a_{ij2}, \ldots, a_{ijs})$ is the vector of design components for the jth subject in the ith group. The other quantities in (33) have been defined earlier. Some components in w_{ij} are shared by other subjects in the same group and thereby generate dependence. Non-shared components produce individual variability in the model. For bivariate times to events, suppressing the group index, the frailty may be expressed in the form (W_1, W_2) where W_k corresponds to the frailty variable for the kth subject, $k = 1, 2$, and is given by

$$W_1 = X_0 + X_1, \text{ and } W_2 = X_0 + X_2,$$

where X_0, X_1, and X_2 are independently distributed positive-valued random variables. The dependence between bivariate times to events arises due to the common term X_0, while the other two terms X_1 and X_2 generate additional unshared variability corresponding to individual random effects. Bayesian inference under this framework has been recently discussed in Mallick and Ravishanker (2006) with application to the tumorigenesis data and to the kidney infection data. Extension to vector frailty is interesting (Xue and Brookmeyer 1996).

Hierarchical frailty models for multilevel multivariate survival data has been discussed by Gustafson (1995) in the context of data from a clinical trial of chemotherapy for advanced colorectal cancer. The data were collected from 419 patients who participated in the trial conducted at 16 clinical sites; Ibrahim *et al.* (2001) have described the use of the hybrid Monte Carlo method for this example. Kuo and Song (2005) have described a dynamic frailty model which assumes a subject's risk changes over time; they used RJMCMC for carrying out inference. Another interesting class of models are multivariate cure rate models (Chen, Ibrahim and Sinha 2002). The cure rate model has been useful for modeling data from cancer clinical trials, where it is assumed that a certain proportion q of the population is cured, while the remaining $1 - q$ is not cured. For bivariate times, see Ibrahim *et al.* (2001, sec 5.5) for details and examples.

6 Disease Mapping

Disease mapping is, broadly speaking, the modeling of the spatial behavior of disease rates, as well as the identification, classification or clustering of areas of highest (lowest) risk rates and their association to explanatory variables. Resource allocation policies and testing of epidemiologic and environmental hypotheses are standard scientific enquiries facilitated by disease mapping models. Statistical models for disease mapping are meaningfully stated as hierarchical models, such as the one characterized by equations (35) to (37) below. The Bayesian approach to disease mapping problems has become commonplace over the last decade, with region-specific random effects modeled by spatially structured priors at one or several hierarchy levels in general hierarchical models (see Section 3.1). It is often useful to present a disease mapping model as a graphical model. This helps us discuss the model structure without distracting details (see Mollié (1996) for an example).

One standard approach in disease mapping is to locally model counts by Poisson distributions, i.e.,

$$y_i|\rho_i \sim \text{Poisson}(\rho_i) \tag{35}$$

where ρ_i is the relative risk in region i, whose structural dependence appears in a second hierarchical level, for instance, as

$$\rho_i = g(x_i'\theta + \beta_i + \epsilon_i) \tag{36}$$

where g is a link function (e.g., exponential), x_i' is a vector of explanatory variables, β_i is a region-specific random effect and ϵ_i is a noise term, usually $N(0, \sigma_\epsilon^2)$. The random effects coefficients β_is follow a standard conditionally autoregressive (CAR) spatial structure (Besag, York and Mollié 1991) with the conditional distribution of β_i depending on β_j for all neighboring regions j, i.e.,

$$\beta_i|\beta_{-i} \sim N\left(\sum_{j\in\delta_i} w_{ij}\beta_j, \sigma_\beta^2/n_i\right) \tag{37}$$

for $\beta_{-i} = (\beta_1, \ldots, \beta_{i-1}, \beta_{i+1}, \ldots, \beta_n)$, δ_i a set of regions adjacent to i, weighting function ω_{ij}, $n_i = \sum_{j\in\delta_i} \omega_{ij}$ and $w_{ij} = \omega_{ij}/n_i$. The most often used neighboring structure is the one that assumes that $\omega_{ij} = 1$ if i and j are neighbor counties and zero otherwise. Bernardinelli, Clayton and Montomoli (1995), Best *et al.* (1999), Kelsal and Wakefield (2002) and Wall (2004) are a few additional studies which use CAR prior distributions for disease mapping.

Example 6 *Nobre, Schmidt and Lopes (2005) examined the spatial and temporal behavior of malaria incidence and its relationship to rainfall over time and across counties, for several counties of Pará, one of Brazil's largest states located in the Amazon region. In 2001, for instance, Pará had around 170 thousand cases of malaria. Malaria affects about 600 million persons a year worldwide and is the most common infectious disease found in Brazil's rainforest. Malaria is transmitted by mosquitoes from the Anopheles sp genus. Temperature and rainfall are important natural risk factors affecting life cycle and breeding of the mosquitoes. Extreme rainfall and extreme drought are both equally likely to lead to proliferation of the mosquitoes and therefore the disease. Limited public health policies and population migration are major social risk factors. Assuming the state of Pará is divided into n contiguous counties (subregions), and that y_i are the number of malaria cases in county i, they extend the standard CAR prior by proposing the following space-time model for malaria counts*

$$\begin{aligned} \rho_{it} &= \exp\{\theta_t x_{it} + \beta_{it}\} \\ \theta_t &\sim N(\theta_{t-1}, \tau_\theta^2) \\ \beta_t &\sim \mathcal{D}(\sigma_t^2) \end{aligned} \tag{38}$$

where x_{it} is a measure of rainfall and $\mathcal{D}$ is the distribution of β_t as a function of σ_t^2. They entertain four space-time models by crossing two specifications for $\mathcal{D}$, i.e., $\beta_t \sim N(0, \sigma_t^2 I_n)$ and $\beta_t \sim CAR(\sigma_t^2)$, and two specifications for σ_t^2, i.e., $\sigma_t^2 \sim IG(a, c)$ and $\sigma_t^2 \sim$ Log-normal$(\log(\sigma_{t-1}^2), \tau^2)$. Figure 1 exhibits log-relative risks (posterior medians) for the month of March 1997 based on the third model specification, i.e. $\beta_t \sim N(0, \sigma_t^2 I_n)$ and $\sigma_t^2 \sim$ Log-normal$(\log(\sigma_{t-1}^2), \tau^2)$. The temporal structures resemble dynamic linear models and dynamic generalized linear models (West, Harrison and Migon 1985).

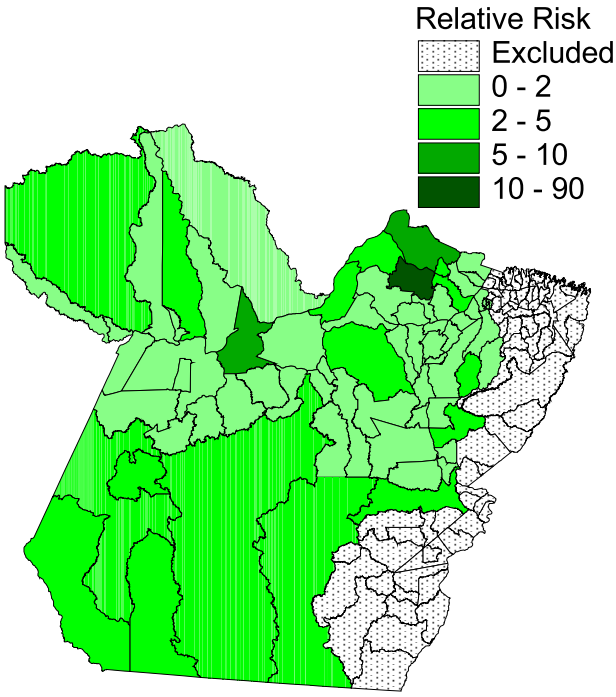

Figure 1: Malaria in Pará: log-relative risk's March, 1997 posterior median when $\beta_t \sim N(0, \sigma_t^2 I_n)$ and $\sigma_t^2 \sim$ Log-normal$(\log(\sigma_{t-1}^2), \tau^2)$. For instance, Anajás county highest risk may be due to its proximity to several rivers and to the island of Marajó.

Nobre, Schmidt and Lopes (2005) generalize the models that appear in Waller *et al.* (1997) and Knorr-Held and Besag (1998), who analyze lung cancer mortality in the state of Ohio over the years. Additional recent space-time studies in disease mapping are Assunção, Reis and Oliveira (2001) who model the diffusion and prediction of Leishmaniasis and Knorr-Held and Richardson (2003) who examine surveillance data on meningococcal disease incidence. MacNab *et al.* (2004), Congdon (2005, Chapter 8), Best, Richardson and Thomson (2005) and references therein provide additional discussion about the Bayesian and empirical Bayesian estimation in disease mapping.

7 Bayesian Clinical Trials

Berry (2006) argues that a Bayesian approach is natural for clinical trial design and drug development. An important advantage is that the Bayesian approach allows for gradual updates of knowledge, rather than restricting the process to updating in large discrete steps measured in trials or phases. The process of updating information under a Bayesian approach is "specifically tied to decision making, within a particular trial, within a drug development program, and within

establishing a company's portfolio of drugs under development" (Berry 2006). He argues that therapeutic areas in which the clinical endpoints are observed early should benefit most from a Bayesian approach. Bayesian methods are particularly useful for statistical inference related to diseases such as cancer in which there is a burgeoning number of biomarkers available for assessing the disease's progress (Berry 2006). An example is presented of a Phase II neoadjuvant HER2/neu-positive breast cancer trial conducted at M. D. Anderson Cancer Center, in which 164 patients were randomized to two treatment arms, chemotherapy with and without trastuzumab (Herceptin; Genentech), and where the primary endpoint was pathological complete response (pCR) of the tumor. In the middle of the trial designed from a frequentist perspective and with the protocol specifying no interim analyses, data available to assess pCR on 34 patients showed that the trastuzumab arm had dramatic improvement, i.e., 4 of 16 control patients (25%) and 12 of 18 trastuzumab patients (67%) experienced a pCR. The Bayesian predictive probability of the standard frequentist statistical significance when 164 patients had been treated was computed to be 95%, demonstrating the use of Bayesian analysis in conjunction with a frequentist design to override the protocol and stop the trial.

In this section we discuss principles of Bayesian clinical trial design, how it relates to frequentist design, and we explain details of some typical Bayesian clinical trial designs.

7.1 Principles of Bayesian Clinical Trial Design

The planning of a clinical trial involves many unknown quantities, including patient responses, i.e., future data, as well as unknown parameters that are never observed. Un certainties about these quantities are best described by defining appropriate probability models. Probability models that are defined on observable data as well as parameters are known as Bayesian models. Clinical trial designs based on such probability models are referred to as Bayesian clinical trial designs. Augmenting the Bayesian model for data and parameters with a formal description of the desired decision leads to a Bayesian decision problem. We refer to clinical trial designs based on this setup as Bayesian decision theoretic designs. Many popular designs stop short of a decision theoretic formulation. Designs that are based on a Bayesian probability model, without a formal definition of a loss function are referred to as proper Bayesian designs (Spiegelhalter, Abrams and Myles 2004).

Typical examples of proper Bayesian approaches are Thall, Simon and Estey (1995) or Thall and Russell (1998).

Example 7 *Consider the following stylized example for an early phase trial. We assume that the observed outcome is an indicator $y_i \in \{0, 1\}$ for tumor response for the i-th patient. Let $\theta = Pr(y_i = 1)$ denote the unknown probability of tumor response. Assume that the current standard of care for the specific disease has a known success probability of $\theta_0 = 15\%$. Let n denote the number of currently enrolled patients, and let $x = \sum_{i=1}^{n} y_i$ denote the recorded number of tumor responses. Assuming a Beta prior, $\theta \sim Be(a, b)$ and a binomial sampling model, $x \sim Bin(n, \theta)$, we can at any time evaluate the posterior distribution $p(\theta \mid y_1, \ldots, y_n) = Be(a + x, b + n - x)$.*

A typical proper Bayesian design could proceed with the following protocol.

- *After each patient cohort, update the posterior distribution.*
- *Stop and recommend the experimental therapy if $Pr(\theta > 0.2 \mid y_1, \ldots, y_n) > \pi_1$.*
- *Stop and abandon the experimental therapy if $Pr(\theta < 0.1 \mid y_1, \ldots, y_n) > \pi_2$ or $n > n_{max}$*

The design requires the elicitation of the prior parameters (a, b) and the choice of policy parameters (tuning parameters) (π_1, π_2, n_{max}). Policy parameters are determined by matching desired operating characteristics, as shown below.

7.2 Operating Characteristics

The distinction between Bayesian and classical (or frequentist) design is not clear cut. Any Bayesian design can be considered and evaluated from a frequentist perspective. As before, let θ and y generically denote the parameters in the underlying probability model and the observed outcomes, respectively. Let δ denote the design. In particular, δ might include a rule to allocate patients to alternative treatment arms, a stopping rule and a final decision. In Example 1, the design $\delta = (d, a)$ included a stopping rule $d(y) \in \{0, 1\}$ with $d = 1$ indicating stopping, and a final decision $a(y) \in \{0, 1\}$ with $a = 1$ indicating a recommendation of the experimental therapy. The recommendation might imply a decision to launch a following confirmatory trial.

For a given set of policy parameters and an assumed truth, we can evaluate frequentist error rates and other properties by evaluating repeated sampling expectations of the relevant summaries. Formally, we consider $E(g(d, \theta, y) \mid \theta)$ for an assumed truth θ. The choice of g and θ depends on the desired summary. For example, Type-I error is evaluated by setting $\theta = \theta_0$ and using $g(d, \theta, y) = I(a = 1)$. To evaluate power we would consider $g(d, \theta, y) = I(a = 1)$ for a grid of θ values with $\theta > 0.2$. Other important summaries are the expected sample size, $g(d, \theta, y) = \min_{n=1,2,\ldots,n_max}\{n : \ d(y_1, \ldots, y_n) = 1\}$ and the expected number of successfully treated patients $g(d, \theta, y) = \sum y_i$. Such summaries are routinely reported as operating characteristics of a design. Formally, these summaries are evaluated by essentially ignoring the Bayesian nature of the design, and considering it as a possible frequentist design. The use of Bayes rules to construct promising candidates for a good frequentist procedure is a commonly used approach, even beyond the context of clinical trial design. It is considered good practice to report operating characteristics when proposing a Bayesian design. In most regulatory or review settings such reports are mandatory.

Besides the reporting purpose, an important use of operating characteristics is to select policy parameters. Like most clinical trial designs, many Bayesian designs require the selection of various policy parameters, such as (π_1, π_2, n_{max}) in the earlier example. A commonly used procedure is to evaluate operating characteristics for a variety of choices of the policy parameters and fixing the final design by matching desired values for the operating characteristics. The resulting design is valid as both, a bona fide frequentist procedure as well as a coherent Bayesian design.

7.3 A Two-Agent Dose-Finding Design

Extensive recent reviews of Bayesian designs for early phase trials appear in Estey and Thall (2003), Berry (2005a) or Berry (2005b). A more comprehensive review, including issues beyond clinical trial design, is presented by Spiegelhalter, Abrams and Myles (2004).

As a typical example for a non-trivial Bayesian design we review in this section a design proposed in Thall *et al.* (2003). They consider a protocol for dose-finding with two agents. The response is an indicator for toxicity, $y_i \in \{0, 1\}$, for the $i-$th patient. For given doses $x = (x_1, x_2)$ of the two agents, let $\pi(x_1, x_2, \theta)$ denote the probability of toxicity. We assume

$$\pi(x, \theta) = \frac{\alpha_1 x_1^{\beta_1} + \alpha_2 x_2^{\beta_2} + \alpha_3 \left(x_1^{\beta_1} x_2^{\beta_2}\right)^{\beta_3}}{1 + \alpha_1 x_1^{\beta_1} + \alpha_2 x_2^{\beta_2} + \alpha_3 \left(x_1^{\beta_1} x_2^{\beta_2}\right)^{\beta_3}}.$$

Here $\theta = (\alpha_1, \beta_1, \alpha_2, \beta_2, \alpha_3, \beta_3)$. The model is chosen to imply parsimonious submodels for the corresponding single agent therapies with $x_1 = 0$ and $x_2 = 0$, respectively. This allows us to include available substantial prior information for (α_1, β_1) and (α_2, β_2). In the application the two agents are gemcitibine and cyclophosphamide, two chemotherapy agents that are extensively used in the treatment of various cancers. Without loss of generality, assume that both agents are available at doses $x_j \in \{0, 0.1, \ldots, 0.9, 1.0\}$.

The proposed design proceeds in two stages. In a first stage we escalate the dose of both agents along a pre-defined grid $D_1 = \{x^1, \ldots, x^k\}$ of dose pairs (x_1^j, x_2^j). For example, the pre-defined grid could be $(x_1^j, x_2^j) = (j \cdot 0.1,\, j \cdot 0.1)$, $j = 1, \ldots, 8$. Patients are assigned in cohorts of K patients, using for example, $K = 2$. Let $Y_i = (y_1, \ldots, y_i)$ denote the recorded responses of the first i patients. After each patient cohort we evaluate the posterior distribution $p(\theta \mid Y)$ for all i currently enrolled patients. The posterior on θ implies a posterior estimated toxicity surface $E(\pi(\theta, x) \mid Y_i)$. Subject to an overdose control, patients in the next cohort are assigned to the dose combination x^j that is closest to a desired target level of toxicity π^*. Overdose control means that no dose combination x^j on the grid can be skipped, i.e., patients can only be assigned to x^j when earlier patients were earlier assigned to x^{j-1}.

After a predetermined number of patients, say n_1, the design switches to the second stage. In the second stage, we drop the restriction to the grid D_1. In words, the assumption is that stage one has approximately identified a dose combination (x_1^*, x_2^*) on the grid D_1 with the desired toxicity, and we can now vary the doses x_1 and x_2 of the two agents to optimize cancer killing and learning about the toxicity surface. This optimization is carried out among all dose pairs with the same a posteriori estimated toxicity level, $E(\pi(x_1, x_2, \theta) \mid Y) \approx \pi^*$. Here Y are all responses that were observed up to now. Cancer killing is approximated as $\lambda(x_1 - x_1^*) + (x_2 - x_2^*)$. This is based on the assumption that the cancer-killing effect of agent 1 is stronger than that of agent 2 by a factor λ, and that the cancer killing effects of both agents is additive and proportional to the dose. Learning about θ is formalized as the log of the determinant of the Fisher information matrix.

In summary, the design has several policy parameters, the choice of the stage one grid D_1, the sample size n_1, the cohort size K, and other parameters that control details that we did not describe above. All policy parameters are chosen by matching desired frequentist operating characteristics.

8 Microarray Data Analysis

8.1 Introduction

High-throughput assays like microarrays, serial analysis of gene expression (SAGE), and protein mass spectrometry are becoming increasingly more commonly used in medical and biological research. Microarrays and SAGE experiments measure mRNA expression, while mass/charge spectra record protein abundance. An excellent review of the experimental setup for all three formats, and related statistical methods appears in Baggerly, Coombes and Morris (2006). See also Datta *et al.* (2007) and Wagner and Naik (2007) in this volume.

Microarray experiments are by far the most commonly used of the three mentioned high-throughput assays. Many published methods for statistical inference (after completed pre-processing) focus on the stylized setup of two-group comparisons. Two-group comparisons are experiments that record gene expression for samples under two biologic conditions of interest and seek to identify those genes that are differentially expressed. Most microarray experiments involve more complicated designs. But the two-group comparison serves as a good canonical example. Popular methods include the use of two-sample t-tests and non-parametric permutation tests, applied to one gene at a time. Several methods have been proposed to adjust the resulting p-values for multiplicities and compile a list of differentially expressed genes. This includes popular methods based on controlling (frequentist) expectation of the false discovery rate (FDR) (Benjamini and Hochberg 1995; Storey 2002; and Storey, Taylor and Siegmund 2004), the beta-uniform method (Pounds and Morris 2003), or SAM (Tusher, Tibshirani and Chu 2001). Here, FDR is defined as the number of false discoveries, that is, the number of genes that are falsely reported as differentially expressed, relative to the total number of genes that are reported as differentially

expressed. The beta-uniform method is based on modeling the distribution of p-values across all genes, and SAM is an algorithm that uses repeated simulation to determine significance cutoffs.

In this section we review some Bayesian inference for group comparison microarray experiments and a Bayesian perspective to error control in massive multiple comparisons. One of the more popular methods is the empirical Bayes approach proposed in Efron *et al.* (2001). The authors describe their approach as an empirical Bayes method. They assume that data are summarized as a set of difference scores, with one score for each gene. The method assumes that these scores arise from a mixture model with submodels corresponding to differentially and non-differentially expressed genes. The desired inference of identifying differentially expressed genes is formally described as the problem of deconvolution of this mixture. This is achieved by clever, but ad-hoc methods. Parmigiani *et al.* (2002) assume a mixture model with three terms corresponding to over-, under- and normal-expressed genes, using uniform distributions for over- and under-expression, and a central Gaussian for normal expression. The authors argue that further inference should be based on the imputed latent trinary indicators in this mixture. Ibrahim, Chen and Gray (2002) propose a model with an explicit threshold to accommodate the many genes that are not strongly expressed, and proceed then with a mixture model including a point mass for the majority of not expressed genes and a log normal distribution for expressed genes. Another class of methods is based on a Gamma/Gamma hierarchical model developed in Newton *et al.* (2001). The model includes parameters that are interpreted as latent indicators for differential expression. Other recent proposals based on mixture models and model selection include Ghosh (2004), Ishwaran and Rao (2003), Tadesse, Ibrahim and Mutter (2003), Broet, Richardson and Radvanyi (2002), Dahl (2003), Tadesse, Sha and Vannucci (2005), Hein *et al.* (2005) and Lewin *et al.* (2006) develop approaches based on hierarchical models. Frigessi *et al.* (2005) develop a hierarchical model based on a detailed description of the physical process, including the details of hybridization, etc.

In this section we review three of the mentioned approaches as typical examples for this literature.

8.2 The Gamma/Gamma Hierarchical Model

In general, a model for microarray group comparison experiments should be sufficiently structured and detailed to reflect the prior expected levels of noise, a reasonable subjective judgment about the likely numbers of differentially expressed genes, and some assumption about dependencies, if relevant. It should also be easy to include prior data when available. One model that achieves these desiderata is the introduced in Newton *et al.* (2001) and Newton and Kendziorski (2003).

The model does not include details of the data cleaning process, including spatial dependence of measurement errors across the microarray, correction for misalignments, etc. While such details are important, it is difficult to automate inference in the form of a generally applicable probability model. We feel normalization and standardization are best dealt with on a case by case basis, exploiting available information about the experimental setup. The remaining variability resulting from preprocessing and normalization can be subsumed as an aggregate in the prior description of the expected noise. So in the following discussion we assume that the data are appropriately standardized and normalized and that the noise distribution implicitly includes these considerations. See, for example, Tseng *et al.* (2001), Baggerly *et al.* (2001) or Yang *et al.* (2002) for a discussion of the process of normalization.

We focus on the comparison of two conditions and assume that data will be available as arrays of appropriately normalized intensity measurements X_{ij} and Y_{ij} for gene i, $i = 1, \ldots, n$, and experiment j, $j = 1, \ldots, J$, with X and Y denoting the intensities in the two conditions. Newton et al. (2001) propose probabilistic modeling for the observed gene frequencies in a

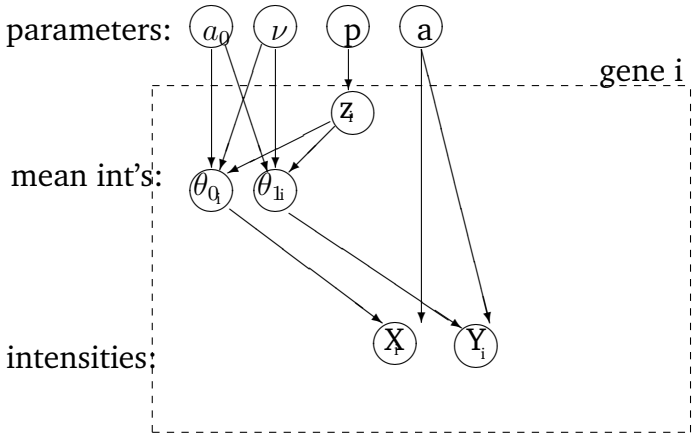

Figure 2: A model for differential gene expression from Newton et al. (2001) for fluorescence intensity measurements in a single slide experiment. For each gene, i.e., each spot on the chip, we record a pair (X, Y) of intensities corresponding to transcript abundance of a gene in both samples. The true unknown mean expression values are characterized by θ_{0i} and θ_{1i}. The aim of the experiment is to derive inference about equality of θ_{1i} and θ_{0k}. Uncertainty about θ_{0i} and θ_{1i} is described by parameters (a_0, ν). The variable z_i is an binomial indicator for equal mean values, i.e., equal expression, associated with a probability p. All information about differential expression of gene i is contained in the posterior distribution for z_i.

single-slide experiment.

For each gene i we record a pair (X_i, Y_i) of intensities corresponding to transcript abundance of a gene in the two samples. The true unknown mean expression levels are denoted by θ_{0i} and θ_{1i}. Other parameters, like scale or shape parameters, are denoted a. The aim of the experiment is to derive inference about the ratio θ_{1i}/θ_{0i}. Uncertainty about θ_{0i} and θ_{1i} is described by hyperparameters (a_0, ν, p). A latent variable $z_i \in \{0, 1\}$ is an indicator for unequal mean values for gene i, i.e., equal expression. We use $z_i = 0$ to indicate equal expression, and $z_i = 1$ to indicate differential expression. Figure 2 summarizes the structure of the probability model. Conditional on the observed fluorescence intensities, the posterior distribution on z_i contains all information about differential expression of gene i. Let $r_i = X_i/Y_i$ denote the observed relative expression for gene i, and let $\eta = (a_0, \nu, p, a)$. Newton et al. (2001) gives the marginal likelihood $p(r_i \mid z_i, \eta)$, marginalized with respect to θ_{1i} and θ_{0i}. They proceed by maximizing the marginal likelihood for η by an implementation of the EM algorithm.

A simple hierarchical extension allows inference for multiple arrays. We assume repeated measurements X_{ij} and Y_{ij}, $j = 1, \ldots, J$, to be conditionally independent given the model parameters. We assume a Gamma sampling distribution for the observed intensities X_{ij}, Y_{ij} for gene i in sample j,

$$X_{ij} \sim \text{Ga}(a, \theta_{0i}) \text{ and } Y_{ij} \sim \text{Ga}(a, \theta_{1i}).$$

The scale parameters are gene specific random effects $(\theta_{0i}, \theta_{1i})$. The model includes an *a priori* positive probability for lack of differential expression

$$Pr(\theta_{0i} = \theta_{1i}) = Pr(z_i = 0) = p.$$

Conditional on latent indicators z_i for differential gene expression, $z_i = I(\theta_{0i} \neq \theta_{1i})$, we assume conjugate gamma random effects distributions

$$\begin{aligned} \theta_{0i} &\sim \text{Ga}(a_0, \nu) \\ (\theta_{1i} | z_i = 1) &\sim \text{Ga}(a_0, \nu) \text{ and } (\theta_{1i} | z_i = 0) \sim \mathbb{I}_{\theta_{0i}}(\theta_{1i}). \end{aligned} \tag{39}$$

The model is completed with a prior for the parameters $(a, a_0, \nu, p) \sim \pi(a, a_0, \nu, p)$ We fix ν, assume *a priori* independence and use marginal gamma priors for a_0 and a, and a conjugate beta prior for p.

Let $X_i = (X_{ij},\ j = 1, \ldots, J)$ and $Y_i = (Y_{ij},\ j = 1, \ldots, J)$. The above model leads to a closed form marginal likelihood $p(X_i, Y_i|\eta)$ after integrating out θ_{1i}, θ_{0i}, but still conditional on $\eta = (p, a, a_0)$.

Availability of the closed form expression for the marginal likelihood greatly simplifies posterior simulation. Marginalizing with respect to the random effects reduces the model to the 3-dimensional marginal posterior $p(\eta \mid y) \propto p(\eta) \prod_i p(X_i, Y_i|\eta)$. Conditional on currently imputed values for η we can at any time augment the parameter vector by generating $z_i \sim p(z_i \mid \eta, X_i, Y_i)$ as simple independent Bernoulli draws, if desired. This greatly simplifies posterior simulation.

8.3 A Non-parametric Bayesian model for Differential Gene Expression

Do, Müller and Vannucci (2005) proposed a semi-parametric Bayesian approach to inference for microarray group comparison experiments. Following the setup in Efron *et al.* (2001), they assume that the data are available as a difference score z_g for each gene, $g = 1, \ldots, G$. For example, the difference score z_g could be a two-sample t-statistic computed with the measurements recorded for gene g on all arrays, arranged in two groups by biologic condition. The summary is a t-statistic only in name, i.e., we do not assume that the sampling model for the statistic is t-distribution under the null hypothesis. Instead, inference proceeds by assuming that the difference scores z_g arise by independent sampling from some unknown distribution f_1 for differentially expressed genes; and from an unknown distribution f_0 for non-differentially expressed genes. For a reasonable choice of difference scores, the distribution f_0 should be a unimodal distribution centered at zero. The distribution f_1 should be bimodal with symmetric modes to the left and right of zero corresponding to over- and under-expressed genes. Of course, the partition into differentially and non-differentially expressed genes is unknown. Thus, instead of samples from f_0 and f_1, we can only work with the samples generated from a mixture $f(z) = p_0\, f_0(z) + (1 - p_0)\, f_1(z)$. Here p_0 is an unknown mixture weight. The desired inference about differential expression for each gene is formalized as a deconvolution of this mixture. We proceed by defining prior probability models on the unknown f_0 and f_1. Probability models on random distributions are traditionally known as nonparametric Bayesian models. See, for example, Müller and Quintana (2004) for a review of non-parametric Bayesian methods. We argue that the marginal posterior probability of gene g being differentially expressed can be evaluated as posterior expectation of $P_g \equiv (1 - p_0) f_1(z_g)/f(z_g)$.

8.4 The Probability of Expression (POE) Model

The POE model is described in Parmigiani *et al.* (2002). A key features of the model is the use of trinary indicators for over- and under-expression. In particular, let y_{ij} denote the observed gene expression for gene i in sample j, with $i = 1, \ldots, n$ and $j = 1, \ldots, J$. We introduce a latent variable $e_{ij} \in \{-1, 0, 1\}$ and assume the following mixture of normal and uniform model, parameterized with $\theta_S = (a_j, \mu_i, s_i, \kappa_i^-, \kappa_i^+)$:

$$p(y_{ij} \mid e_{ij}) \sim f_{eg}(y_{ij}) \text{ with } \left\{ \begin{array}{rcl} f_{-1i} & = & U(-\kappa_i^- + a_j + \mu_i, a_j + \mu_i) \\ f_{0i} & = & N(\quad a_j + \mu_i, s_i) \\ f_{1i} & = & U(\quad a_j + \mu_i, a_j + \mu_i + \kappa_i^+). \end{array} \right. \tag{40}$$

In other words, we assume that the observed gene expressions arise from a mixture of a normal distribution and two uniform distributions defined to model over-dispersion relative to the normal. Conditional on the parameters and the latent indicators e_{ij}, we assume that the observed

gene expressions y_{ij} are independent across genes and samples. The interpretation of the normal component is as a baseline distribution for gene i, and the two uniform terms as the distribution in samples where gene i is over- and under-expressed, respectively. In (40), a_j, $j = 1, \ldots, J$ are sample specific effects, allowing inference to adjust for systematic variation across samples, μ_i are gene specific effects that model the overall prevalence of each gene, and κ_i^- and κ_i^+ parameterize the over-dispersion in the tails. Finally, s_i is the variance of the baseline distribution for gene i. Parmigiani *et al.* (2002) define (conditionally) conjugate priors for μ_i, s_i and κ_i^+ and κ_i^-. For the slide specific effect we impose an identifiability constraint $a_j \sim N(0, \tau^2)$, i.i.d., subject to $\sum a_j = 0$.

8.5 Multiplicity Correction: Controlling False Discovery Rate

High throughput gene expression experiments often give rise to massive multiple comparison problems. We discuss related issues in the context of microarray group comparison experiments. Assume that for genes, $i = 1, \ldots, n$, for large n, we wish to identify those that are differentially expressed across two biologic conditions of interest. ¿From a classical perspective, multiple comparisons require an adjustment of the error rate, or, equivalently, an adjustment of the nominal significance level for each comparison. This is achieved, for example, in the Bonferroni correction or by the Benjamini and Hochberg (1995) correction mentioned in Section 8.1.

It can be argued that Bayesian posterior inference already accounts for multiplicities, and no further adjustment is required (Scott and Berger 2006). The argument is valid for the evaluation of posterior probabilities of differential expression. In a hierarchical model, the reported posterior probabilities are correctly adjusted for the multiplicities. But reporting posterior probabilities only solves half the problem. We still need to address the second step of the inference problem, namely the identification of differentially expressed genes. Berry and Hochberg (1999) discuss this perspective.

This identification is most naturally discussed as a decision problem. The formal statement of a decision problem requires a probability model $p(\theta, y)$ for all unknown quantities, including parameters θ and data y, a set of possible actions, $\delta \in D$, and a loss function $L(\delta, \theta, y)$ that formalizes the relative preferences for decision δ under hypothetical outcomes y and assumed parameter values θ. The probability model could be, for example, the hierarchical gamma/gamma model described in Section 8.2. The action is a vector of indicators, $\delta = (\delta_1, \ldots, \delta_n)$ with $\delta_i = 1$ indicating that the gene is reported as differentially expressed. We write $\delta(y)$ when we want to highlight the nature of δ as a function of the data. Let $r_i \in \{0, 1\}$ denote an indicator for the (unknown) truth. The r_i are part of the parameter vector θ. Usually, the probability model includes additional parameters besides r. It can be argued (Robert 2001) that a rational decision maker should select the rule $\delta(y)$ that maximizes L in expectation. The relevant expectation is the probability model on θ conditional on the observed data, leading to the optimal rule

$$\delta^*(y) = \arg\min_{\delta} \int L(\delta, \theta, y)\; p(\theta \mid y)\; \mathrm{d}\theta.$$

Let $v_i = E(r_i \mid y)$ denote the marginal posterior probability of gene i being differentially expressed. The assumption of non-zero prior probabilities, $0 < p(r_i = 1) < 1$, ensures non-trivial posterior probabilities. In Müller et al. (2004) we show that for several reasonable choices of $L(\delta, \theta, y)$ the optimal rule is of the form $\delta_i^*(y) = I(v_i > t)$. In words, the optimal decision is to report all genes with marginal probability of differential expression beyond a certain threshold t as differentially expressed. The value of the threshold depends on the specific loss function. The optimal rule δ^* is valid for several loss functions defined in Müller *et al.* (2004). Essentially, all are variations of basic 0-1 loss functions. Let

$$\text{FD} = \sum \delta_i\,(1 - r_i) \quad \text{and} \quad \text{FN} = \sum (1 - \delta_i)\, r_i$$

denote false discovery and negative counts, and let

$$\text{FDR} = \text{FD}/\sum \delta_i \quad \text{and} \quad \text{FNR} = \text{FN}/\sum (1-\delta_i)$$

denote false discovery and false negative rates. The definitions FD(R) and FN(R) are summaries of parameters, r, and data, $\delta(y)$. Taking an expectation with respect to y and conditioning on r one would arrive at the usual definition of false discovery rates, as used, among many others, in Benjamini and Hochberg (1995), Efron and Tibshirani (2002), Storey (2002, 2003), and Storey, Taylor and Siegmund (2004). Instead we use posterior expectations, defining $\overline{\text{FD}} = E(\text{FD} \mid y)$, etc. See, Genovese and Wasserman (2002, 2004) for a discussion of posterior expected FDR. Using these posterior summaries we define the following losses: $L_N(\delta, z) = c\overline{\text{FD}} + \overline{\text{FN}}$, and $L_R(\delta, z) = c\,\overline{\text{FDR}} + \overline{\text{FNR}}$. The loss function L_N is a natural extension of $(0, 1, c)$ loss functions for traditional hypothesis testing problems (Lindley 1971). Alternatively, we consider bivariate loss functions that explicitly acknowledge the two competing goals: $L_{2R}(\delta, z) = \overline{\text{FNR}}$, subject to $\overline{\text{FDR}} < \alpha_R$, and $L_{2N}(\delta, z) = \overline{\text{FN}}$, subject to $\overline{\text{FD}} < \alpha_N$. Under all four loss functions, L_N, L_R, L_{2R} and L_{2N}, the nature of the optimal rule is δ^*. See Müller et al. (2004) for the definition of the thresholds.

One can argue that not all false negatives and all discoveries are equally important. False negatives of genes that are massively differentially expressed are more serious than only marginally differentially expressed genes. To formalize this notion we need to assume that the probability model includes parameters that can be interpreted as extent of differential expression, or strength of the signal. Assume that the model includes parameters m_i, $i = 1, \ldots, n$, with $m_i > 0$ if $r_i = 1$ and $m_i = 0$ if $r_i = 0$. For example, in the gamma/gamma hierarchical model of Section 8.2 a reasonable definition would use $m_i = \log(\theta_{i1}/\theta_{i0})$. Assuming parameters m_i that can be interpreted as level of differential expression for gene i, we define

$$L_\rho(\rho, \delta, z) = -\sum \delta_i\, m_i + k \sum (1-\delta_i) m_i + c \sum \delta_i.$$

For $c > 0$ the optimal solution is easily found as $\delta_i^* = I\{\bar{m}_i \geq c/(1+k)\}$. For more discussion and alternative loss functions see, for example, Müller, Parmigiani and Rice (2007).

9 Summary

In this chapter we have reviewed the basic framework of Bayesian statistics, and typical inference problems that arise in biomedical applications. In summary, Bayesian inference can be carried out for any problem that is based on a well defined probability framework. In particular, Bayesian inference requires a likelihood, i.e., a sampling model of the observable data conditional on assumed values of the parameters, and a prior, i.e., a prior judgment about the parameters that is formalized as a probability model. As long as the likelihood and the prior are available for any set of assumed parameter values, one can in principle implement Bayesian inference. Evaluation up to a constant is sufficient for MCMC posterior simulation with Metropolis-Hastings chains.

The main advantage of Bayesian inference is the fact that inference is based on a principled and coherent approach. In particular, even for complicated setups with hierarchical models, multiple studies, mixed data types, delayed responses, complicated dependence etc., as long as the investigator is willing to write down a probability model, one can carry out Bayesian inference.

There are some important limitations of the Bayesian approach. We always need a well defined probability model. There are (almost) no genuinely non-parametric Bayesian methods (a class of models known as "non-parametric Bayesian models" are really random functions, i.e., probability models on infinite dimensional spaces). For example, the proportional hazards rate

model for event time data has no easy Bayesian equivalence. Even simple approaches like kernel density estimation, or loess smoothing have no simple Bayesian analogues. Another limitation of Bayesian inference that arises from the need for well defined probability models is the difficulty to define good model validation schemes. Principled Bayesian model comparison is always relative to an assumed alternative model. There is no easy Bayesian equivalence of a simple chi-square test of fit. Although there are several very reasonable Bayesian model validation approaches, none are based on first principles. Another great limitation of Bayesian inference is the sensitivity to the prior probability model. This is usually not a problem when the goal of the data analysis is prediction or parameter estimation. Results about posterior asymptotics assure us that the impact of the prior choice will eventually wash out. However, the same is not true for model comparison. Bayes factors are notoriously sensitive to prior assumptions, and there is no easy way to avoid this.

In summary, the increasing notion that the advantages outweigh the limitations has led to an increasingly greater use of Bayesian methods in several areas of application including biomedical applications. Increasingly more complex inference problems, and increasingly more expensive data collection in experiments or clinical studies requires that we make the most of the available data. For complex designs, Bayesian inference may often be the most feasible way to proceed.

References

Abramowitz, M. and Stegun, I. A. (eds). 1965. *Handbook of Mathematical Functions*, National Bureau of Standards, Washington.

Achcar, J. A., Brookmeyer, R. and Hunter, W. G. 1985. An application of Bayesian analysis to medical follow-up data. *Statistics in Medicine* 4: 509-20.

Albert, J. 1988. Computational methods using a Bayesian hierarchical generalized linear model. *Journal of the American Statistical Association* 83: 1037-44.

Arjas, E. and Gasbarra, D. 1994. Nonparametric Bayesian inference for right-censored survival data, using the Gibbs sampler. *Statistica Sinica* 4: 505-24.

Assunção, R. M. and Reis, I. A. and Oliveira, C. D. L. 2001. Diffusion and prediction of Leishmaniasis in a large metropolitan area in Brazil with a Bayesian space-time model. *Statistics in Medicine* 20: 2319-35.

Baggerly, K. A., Coombes, K. R., Hess, K. R., Stivers, D. N., Abruzzo, L. V. and W., Z. 2001. Identifying differentially expressed genes in cDNA microarray experiments. *Journal Computational Biology*, 8: 639-59.

Baggerly, K. A., Coombes, K. R. and Morris, J. S. 2006. An introduction to high-throughput bioinformatics data. In *Bayesian Inference for Gene Expression and Proteomics*, eds. K.-A. Do, P. Müller and M. Vannucci), 1-34. Cambridge: Cambridge University Press.

Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, B* 57: 289-300.

Berger, J. O. 1985. *Statistical Decision Theory and Bayesian Analysis*. Berlin: Springer-Verlag.

Berger, J. O. and Bernardo, J. M. 1992. On the development of reference priors. In *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, 35-60. Oxford: Oxford University Press.

Berger, J. O. and Pericchi, L. R. 1996. The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association* 91: 109-22.

Berger, J. O., Insua, D. R. and Ruggeri, F. 2000. Bayesian robustness. In *Robust Bayesian Analysis*, eds. D. R. Insua and F. Ruggeri, 1-32, New York: Spriger-Verlag.

Bernardinelli, L. and Clayton, D. and Montomoli, C. 1995. Bayesian estimates of disease maps: How important are priors?. *Statistics in Medicine* 14: 2411-2431.

Bernardo, J. M. 1979. Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc., Ser. B* 41: 113-47.

Bernardo, J. M. and Smith, A. F. M. 1994. *Bayesian Threory*. New York: Wiley.

Berry, D. 2005(a). The Bayesian principle: can we adapt? yes! *Stroke* 36: 1623-4.

— 2005(b). Clinical trials: Is the Bayesian approach ready for prime time? yes! *Stroke* 36: 1621-2 [commentary].

— 2006. A guide to drug discovery: Bayesian clinical trials. *Nature Reviews Drug Discovery* 5: 27-36.

Berry, D. A. and Hochberg, Y. 1999. Bayesian perspectives on multiple comparisons. *Journal of Statistical Planning and Inference* 82: 215-27.

Besag, J., York, J. and Mollié, A. 1991. Bayesian image restoration, with applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics* 43: 1-59.

Best, N. G., Arnold, R. A., Thomas, A., Waller, L. A. and Conlon, E. M. 1999. Bayesian models for spatially correlated disease and exposure Data. In *Bayesian Statistics 6*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, 131-56. Oxford: Oxford University Press.

Best, N. G., Richardson, S. and Thomson, A. 2005. A comparison of Bayesian spatial for disease mapping. *Statistical Methods in Medical Research*, **14**: 35-59.

Best, N. and Thomas, A. 2000. Bayesian graphical models and software for GLM's. In *Generalized Linear Models: A Bayesian Perspective*, eds. D. K. Dey, S. K. Ghosh and B. K. Mallick, 387-402. New York: Marcel Dekker.

Broet, P., Richardson, S. and Radvanyi, F. 2002. Bayesian hierarchical model for identifying changes in gene expression from microarray experiments. *J Comput Biol.* 9: 671-83.

Buckle, D. J. 1995. Bayesian inference for stable distributions. *Journal of the American Statistical Association* 90: 605-613.

Buonaccorsi, J. P. and Gatsonis, C.A. 1988. Bayesian inference for ratios of coefficients in a linear model. *Biometrics* 44: 87-101.

Cappé O., Guillin, A., Marin, J. M., and Robert, C. P. 2004. Population Monte Carlo. *J. Comput. Graph. Stat.* 13: 907-29.

Carlin, B. P. and Chib, S. 1995. Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B* 57: 473-84.

Carlin, B. P. and Hodges, J. S. 1999. Hierarchical proportional hazards regression models for highly stratified data. *Biometrics* 55: 1162-70.

Carlin, B. P. and Louis, T. 2000. *Bayes and Empirical Bayes Methods for Data Analysis*. London: Chapman & Hall/CRC.

Chen, H.-H., Shao, Q. M. and Ibrahim, J. G. 2000. *Monte Carlo Methods in Bayesian Computation*. New York: Springer- Verlag.

Chen, M.-H. and Dey, D. K. 2000. Bayesian analysis for correlated ordinal data models. In *Generalized Linear Models: A Bayesian Perspective*, eds. D. K. Dey, S. K. Ghosh and B. K. Mallick, 133-55. New York: Marcel Dekker.

Chen, M.-H., Ibrahim, J. G. and Sinha, D. 2002. Bayesian inference for multivariate survival data with a surviving fraction. *Journal of Multivariate Analysis* 80: 1011-26.

Clayton, D. and Cuzick, J. 1985. Multivariate generalizations of the proportional hazards model. *Journal of the Royal Statistical Society A* 148: 82-117.

Clayton, D. and Kaldor, J. 1987. Empirical Bayes estimates of age-standardised relative risks for use in disease mapping. *Biometrics* 43: 671-81.

Clayton, D. 1996. Generalized linear mixed models. In *Markov Chain Monte Carlo in Practice*, eds. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, 275-301, New York: Chapman & Hall.

Clyde, M. A. 1999. Bayesian model averaging and model search strategies. In *Bayesian Statistics 6*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, 157-85. Oxford: Oxford University Press.

Congdon, P. 2003. *Applied Bayesian Modeling*. New York: Wiley.

Congdon, P. 2005. *Bayesian Models for Categorical Data*. New York: Wiley.

Dahl, D. 2003. Modeling differential gene expression using a Dirichlet process mixture model. In *2003 Proceedings of the American Statistical Association, Bayesian Statistical Sciences Section*. Alexandria, VA: American Statistical Association.

Datta, S., Datta, S., Parresh, R. S. and Thompson, C. M. 2007. Microarray data analysis. In *Computational Methods in Biomedical Research*, eds. R. Khattree and D. N. Naik, this volume, pp.

Dellaportas, P. and Smith, A. F. M. 1993. Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling. *Applied Statistics* 42: 443-59.

Dellaportas, P., Forster, J. J. and Ntzoufras, I. 2000. Bayesian variable selection using the Gibbs sampler. In *Generalized Linear Models: A Bayesian Perspective*, eds. D. K. Dey, S. K. Ghosh and B. K. Mallick, 273-83. New York: Marcel Dekker.

Dellaportas, P., Forster, J. and Ntzoufras, I. 2002. On Bayesian model and variable selection using MCMC. *Statistics and Computing* 12: 27-36.

Dempster, A. P., Laird, N. M. and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser B* 39:1-38.

Dey, D. K., Chen, M.-H. and Chang, H. 1997. Bayesian approach for the nonlinear random effects model. *Biometrics* 53: 1239-52.

Dey, D. K. and Ravishanker, N. 2000. Bayesian approaches for overdispersion in generalized linear models. In *Generalized Linear Models: A Bayesian Perspective*, eds. D. K. Dey, S. K. Ghosh and B. K. Mallick, 73-84. New York: Marcel Dekker.

Do, K., Müller, P. and Tang, F. 2005. A Bayesian mixture model for differential gene expression. *Journal of the Royal Statistical Society C* 54: 627-44.

Do, K., Müller, P. and Vannucci, M. 2006. *Bayesian Inference for Gene Expression and Proteomics*. Cambridge: Cambridge University Press.

Doucet, A., de Freitas, N. and Gordon, N. 2001. *Sequential Monte Carlo Methods in Practice*. New York: Springer-Verlag.

Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. 2001. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96: 1151-60.

Estey, E. and Thall, P. 2003. New designs for phase 2 clinical trials. *Blood* 102: 44248.

Faraggi, D. and Simon, R. 1995. A neural network model for survival data. *Statistics in Medicine* 14: 73-82.

Frigessi, A., van de Wiel, M., Holden, M., Glad, I., Svendsrud, D. and Lyng, H. 2005. Genome-wide estimation of transcript concentrations from spotted cDNA microarray data. *Nucleic Acids Res. - Methods Online* 33: e143.

Finkelstein, D. M. and Wolfe, R. A. 1985. A semiparametric model for regression analysis of interval-censored survival data. *Biometrics* 41: 933-45.

Gamerman, D. and Lopes, H. F. 2006. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference (2nd edition)*. Boca Raton: Chapman&Hall/CRC.

Gamerman, D. and Migon, H. S. 1993. Dynamic hierarchical models. *J. Roy. Statist. Soc., B.* 55: 629-42.

Geisser, S. 1975. The predictive sample reuse method with application. *Journal of the American Statistical Association* 70: 320-8, 350.

Gelfand, A. E. and Smith, A. F. M. 1990. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* 85: 398-409.

Gelfand, A. E. and Ghosh, S. 1994. Model choice: a minimum posterior predictive loss approach. *Biometrika* 85: 1-11.

Gelfand, A. E. and Mallick, B. K. 1995. Bayesian analysis of proportional hazards models built from monotone functions. *Biometrics* 51: 843-52.

Gelfand, A. E. and Ghosh, M. 2000. Generalized linear models: a Bayesian view. In *Generalized Linear Models: A Bayesian Perspective*, eds. D. K. Dey, S. K. Ghosh and B. K. Mallick, 3-21. New York: Marcel Dekker.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. 2004. *Bayesian Data Analysis*, New York: Chapman & Hall/CRC.

Geman, S. and Geman, D. 1984. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6: 721-41.

Genovese, C. R. and Wasserman, L. 2002. Operating characteristics and extensions of the FDR procedure. *Journal of the Royal Statistical Society, Series B* 64: 499517.

Genovese, C. R. and Wasserman, L. 2004. A stochastic process approach to false discovery control. *Annals of Statistics* 32: 1035-61.

George, E. I. and McCulloch, R. E. 1992. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 79: 677-83.

Geweke, J. 1989. Bayesian inference in econometric models using Monte Carlo integration. *Econometrica* 57: 1317-39.

Ghosh, D. 2004. Mixture models for assessing differential expression in complex tissues using microarray data. *Bioinformatics* 20: 1663-9.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. 1996 *Markov Chain Monte Carlo in Practice*, Chapman & Hall, London.

Gilks, W. R. and Wild, P. 1992. Adaptive rejection sampling for Gibbs sampling. *Applied Statistics* 41: 337-48.

Godsill, S. J. 2001. On the relationship between Markov chain Monte Carlo methods for model uncertainty. *Journal of Computational and Graphical Statistics* 10: 1-19.

Gordon, K. and Smith, A. F. M. 1990. Modeling and monitoring biomedical time series. *Journal of the American Statistical Association* 85: 328-37.

Green, P. J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82: 711-32.

Gustafson, P. 1996. Robustness considerations in Bayesian analysis. *Stat. Meth. in Medical Res.* 5: 357-73.

Gustafson, P. 1997. Large hierarchical Bayesian analysis of multivariate survival data. *Biometrics* 53: 230-43.

Gustafson, P. 1998. Flexible Bayesian modeling for survival data. *Lifetime Data Analysis* 4: 281-99.

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

Hein, A., Richardson, S., Causton, H., Ambler, G. and Green, P. 2005. Bgx: A fully Bayesian integrated approach to the analysis of affymetrix genechip data. *Biostatistics* 6: 349-73.

Hougaard, P. 2000. *Analysis of Multivariate Survival Data*. New York: Springer-Verlag.

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

Heyde, C. C. and Johnstone, I. M. 1979. On asymptotic posterior normality of stochastic processes. *Journal of the Royal Statistical Society, B* 41: 184-9.

Hoeting, J., Madigan, D., Raftery, A. E. and Volinsky, C. 1999. Bayesian model averaging (with discussion). *Statistical Science* 14: 382-417.

Ibrahim, J. G., Chen, M.-H. and Sinha, D. 2001. *Bayesian Survival Analysis*. New York: Springer-Verlag.

Ibrahim, J., Chen, M.-H. and Gray, R. 2002. Bayesian models for gene expression with DNA microarray data. *Journal of the American Statistical Association*, 97: 88-100.

Ishwaran, H. and Rao, J. S. 2003. Detecting differentially expressed genes in micro- arrays using Bayesian model selection. *Journal of the American Statistical Association* 98: 438-55.

Jeffreys, H. 1961. *Theory of Probability*. Oxford: Oxford University Press.

Kadane, J. B. ed. 1984. *Robustness of Bayesian Analysis*. Amsterdam: North-Holland.

Kalbfleisch, J. D. 1978. Nonparametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society B* 40: 214-21.

Kass, R. E., Tierney, L. and Kadane, J. B. 1988. Asymptotics in Bayesian computation (with discussion). In *Bayesian Statistics 3*, eds. J. M. Bernardo et al.), Oxford: Oxford University Press, 261-78.

Kass, R. E., Tierney, L. and Kadane, J. B. 1989. Approximate methods for assessing influence and sensitivity in Bayesian analysis. *Biometrika* 76: 663-74.

Kass, R. E. and Raftery, A. 1995. Bayes factors and model uncertainty. *Journal of the American Statistical Association* 90: 773-95.

Kass, R. E. and Wasserman, L. 1996. The selection of prior distributions by formal rules. *Journal of the American Statistical Association* 91: 1343-70.

Kelsal, J. and Wakefield, J. 2002. Modeling Spatial Variation in Disease Risk: A Geostatistical Approach, *Journal of the American Statistical Association* 97: 692-701.

Kim, S. W. and Ibrahim, J. G. 2000. On Bayesian inference for proportional hazards models using noninformative priors. *Lifetime Data Analysis* 6: 331-41.

Knorr-Held, L. and Besag, J. 1998. Modelling Risk from a disease in time and space. *Statistics in Medicine* 17: 2045-60.

Knorr-Held, L. and Richardson, S. 2003. A hierarchical model for space-time surveillance data on meningococcal disease incidence. *Applied Statistics* 52: 169-83.

Kristiansen, N. K., Sjöström, S. O. and Nygaard, H. 2005. Urinary bladder volume tracking using a Kalman filter. *Medical and Biological Engineering and Computing* 43: 331-4.

Kuo, L. and Peng, F. 2000. A mixture-model approach to the analysis of survival data. In *Generalized Linear Models: A Bayesian Perspective*, eds. D. K. Dey, S. K. Ghosh and B. K. Mallick, 195-209. New York: Marcel Dekker.

Kuo, L. and Song, C. 2005. A new time varying frailty model for recurrent events. *Proceedings of the ASA section on Bayesian Statistical Science, American Statistical Association*.

Landrum, M. B. and Normand, S. 2000. Developing and applying medical practice guidelines following acute myocardial infarction: a case study using Bayesian probit and logit models. In *Generalized Linear Models: A Bayesian Perspective*, eds. D. K. Dey, S. K. Ghosh and B. K. Mallick, 195-209. New York: Marcel Dekker.

Larget, B. 2005. Introduction to Markov chain monte carlo methods in molecular evolution. In *Statistical Methods in Molecular Evolution*, ed. R. Nielsen, 45-62. New York: Springer-Verlag.

Lavine, M. 1992. Some aspects of Polya-tree distributions for statistical modeling. *Annals of Statistics* 20: 1222-35.

Lee, P. M. 1997. *Bayesian Statistics: An Introduction*. New York: Wiley.

Lewin, A., Richardson, S., Marshall, C., Glazier, A. and Aitman, T. 2006. Bayesian modelling of differential gene expression. *Biometrics*, 62, 1-9.

Lindley, D. V. 1971. *Making decisions*. New York: John Wiley & Sons.

Lindley, D. V. and Smith, A. F. M. 1972. Bayes estimtes for the linear model. *J. Royal Statist. Soc. Ser. B* 34: 1-41.

Lopes, H. F., Müller, P. and Rosner, G. L. 2003. Bayesian meta-analysis for longitudinal data models using multivariate mixture priors. *Biometrics* 59: 66-75.

MacEachern, S. N. and Müller, P. 1998. Estimating mixture of Dirichlet process models. Journal of Computational and Graphical Statistics 7: 223-38.

MacNab, Y. C., Farrell, P. J., Gustafson, P. and Wen, S. 2004. Estimation in Bayesian Disease mapping. *Biometrics* 60: 865-73.

McGilchrist, C. A. and Aisbett, C. W. 1991. Regression with frailty in survival analysis. *Biometrics* 47: 461-6.

Mallick, M. and Ravishanker, N. 2004. Multivariate survival analysis with PVF frailty models. In *Advances in Ranking and Selection, Multiple Comparisons, and Reliability, with Applications*, eds. N. Balakrishnan, N. Kannan and H.N. Nagaraja, 369-84. Boston: Birkhauser.

Mallick, M. and Ravishanker, N. 2006. Additive positive stable frailty models. *Methodology and Computing in Applied Probability* 8: 541-58.

Mantel, N., Bohidar, N. R. and Ciminera, J. L. 1977. Mantel-Haenszel Analyses of Litter-matched Time-to-response Data, with Modifications for Recovery of Interlitter Information. *Cancer Research* 37: 3863-8.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. 1953. Equation of state calculations by fast computing machine. *Journal of Chemical Physics* 21: 1087-91.

Mezzetti, M. and Ibrahim, J. G. 2000. Bayesian inference for the Cox model using correlated gamma process priors. *Technical Report*. Department of Biostatistics, Harvard School of Public Health.

Mollié, A. 1996. Bayesian mapping of disease. In *Markov Chain Monte Carlo in Practice*, eds. W. R. Gilks, S. Richardson and D. J. Spiegelhalter, 359-79. New York: Chapman & Hall.

Müller, P., Parmigiani, G. and Rice, K. 2007. FDR and Bayesian multiple comparisons rules. In *Bayesian Statistics 8*, (eds. J. Bernardo, S. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith and M. West), to appear. Oxford: Oxford University Press.

Müller, P., Parmigiani, G., Robert, C. and Rouseau, J. 2004. Optimal sample size for multiple testing: the case of gene expression microarrays. *Journal of the American Statistical Association* 99: 990-1001.

Müller, P. and Quintana, F. A. 2004. Nonparametric Bayesian data analysis. *Statistical Science* 19: 95-110.

Müller, P., Parmigiani, G., Schildkraut, J. and Tardella, L. 1999. A Bayesian hierarchical approach for combining case-control and prospective studies. *Biometrics* 55: 258-66.

Naylor, J. C. and Smith, A. F. M. 1982. Application of a method for the efficient computation of posterior distributions. *Applied Statistics* 31: 214-25.

Newton, M. A. and Kendziorski, C. M. 2003. Parametric empirical Bayes methods for micorarrays. In *The analysis of gene expression data: methods and software*, (eds. G. Parmigiani, E. S. Garrett, R. Irizarry and S. L. Zeger). New York: Springer Verlag.

Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R. and Tsui, K. W. 2001. On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* 8: 37-52.

Nobre, A. A., Schmidt, A. M. and Lopes, H. F. 2005. Spatio-temporal models for mapping the incidence of malaria in Pará. *Environmetrics* 16: 291-304.

Oakes, D. 1989. Bivariate survival models induced by frailties. *Journal of the American Statistical Association* 84: 487-93.

Parmigiani, G., Garrett, E. S., Anbazhagan, R. and Gabrielson, E. 2002. A statistical framework for expression-based molecular classification in cancer. *Journal of the Royal Statistical Society B* 64: 717-36.

Pauler, D. K., Wakefield, J. C. and Kass, R. E. 1999. Bayes factors for variance component models. *Journal of the American Statistical Association* 94: 1241-53.

Pole, A., West, M. and Harrison, J. 1994. *Applied Bayesian Forecasting and Time Series Analysis*. Boca Raton: CRC

Pounds, S. and Morris, S. 2003. Estimating thhe occurence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics* 19: 1236-42.

Qiou, Z., Ravishanker, N. and Dey, D. K. 1999. Multivariate survival analysis with positive stable frailties. *Biometrics* 55: 637-44.

R Development Core Team 2006. *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing: Vienna, Austria.

Raftery, A. E., Madigan, D. and Hoeting, J. A. 1997. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* 92: 179-91.

Ravishanker, N. and Dey, D. K. 2000. Multivariate survival models with a mixture of positive stable frailties. *Methodology and Computing in Applied Probability* 2: 293-308.

Robert, C. 2001. *The Bayesian Choice*. New York: Springer-Verlag.

Roberts, C. D. 2000. *Generalized Linear Models: A Bayesian Perspective*. New York: Marcel Dekker.

Rubin, D. B. 1989. A new perspective on meta-analysis. In *The Future of Meta-Analysis*, eds. K. W. Wachter and M. L. Straf. New York: Russell Sage Foundation.

Scott, J. and Berger, J. 2006. An exploration of aspects of Bayesian multiple testing. *Tech. rep.*, Duke University, ISDS.

M.J. Schervish, M. J. 1995. *Theory of Statistics*. New York: Springer-Verlag.

Shumway, R. H. and Stoffer, D. S. 2004. *Time Series Analysis and its Applications*. New York: Springer Verlag.

Sinha, D. and Dey, D. K. 1997. Semiparametric Bayesian analysis of survival data. *Journal of the American Statistical Association* 92: 1195-313.

Sinha, D. 1997. Time discrete Beta-process model for interval censored survival data. *Canadian Journal of Statistics* 25:445-56.

Sinha, D., Chen, M.-H. and Ghosh, S. K. 1999. Bayesian analysis and model selection for interval-censored survival data. *Biometrics* 55: 585-90.

Sisson, S. A. 2005. Transdimensional Markov chains: a decade of progress and future perspectives. *Journal of the American Statistical Association* 100: 1077-89.

Sivaganesan, S. 2000. Global and local robustness: uses and limitations. In *Robust Bayesian Analysis*, eds. D. R. Insua and F. Ruggeri, 89-108, New York: Spriger-Verlag.

Sivia, D. S. 1996. *Data Analysis. A Bayesian Tutorial*. Oxford: Oxford University Press.

Smith, A. F. M. 1973. A general Bayesian linear model. *J. Roy. Statist. Soc. Ser. B* 35: 67-75.

Smith, A. F. M. and Gelfand, A. E. 1992. Bayesian statistics without tears: a sampling-resampling perspective. *American Statistician* 46: 84-8.

Sorensen, D. and Gianola, D. 2002. *Likelihood, Bayesian and MCMC Methods in Quantitative Genetics*. New York: Springer-Verlag.

Spiegelhalter, D. J., Abrams, K. R. and Myles, J. P. 2004. *Bayesian approaches to clinical trails and health care evaluation*. Chichester, UK: John Wiley & Sons.

Spiegelhalter, D. J., Thomas, A. and Best, N. G. 1999. *WinBUGS Version 1.2 User Manual*, Cambridge, UK: MRC Biostatistics Unit.

Storey, J. 2002. A direct approach to false discovery rates. *Journal of the Royal Statistical Society B* 64: 479-98.

Storey, J. D., Taylor, J. E. and Siegmund, D. 2004. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 66: 187-205.

Sun, D., Speckman, P. L. and Tsutakawa, R. K. 2000. Random effects in generalized linear mixed models (GLMM's). In *Generalized Linear Models: A Bayesian Perspective*, eds. D. K. Dey, S. K. Ghosh and B. K. Mallick, 3-36. New York: Marcel Dekker.

Symons, M. J., Grimson, R. C. and Yuan, Y. C. 1983. Clustering of Rare Events. *Biometrics* 39, 193-205.

Tadesse, M., Sha, N. and Vannucci, M. 2005. Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association* 100: 602-17.

Tadesse, M. G., Ibrahim, J. G. and Mutter, G. L. 2003. Identification of differen- tially expressed genes in high-density oligonucleotide arrays accounting for the quantification limits of the technology. *Biometrics* 59: 542-54.

Tanner, M. A. and Wong, W. 1987. The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association* 82: 528-50.

Thall, P. F., Millikan, R. E., Müller, P. and Lee, S.-J. 2003. Dose-finding with two agents in phase I oncology trials. *Biometrics* 59: 487-96.

Thall, P. F. and Russell, K. E. 1998. A strategy for dose-finding and safety monitoring based on efficacy and adverse outcomes in phase I/II clinical trials. *Biometrics* 54: 251-64.

Thall, P. F., Simon, R. M. and Estey, E. H. 1995. Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. *Statistics in Medicine* 14: 357-79.

Tierney, L. 1994. Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics* 22: 1701-62.

Tierney, L. and Kadane, J. B. 1986. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association* 81: 82-6.

Tierney, L., Kass, R. E. and Kadane, J. B. 1989. Fully exponential Laplace approximations for expectations and variances of nonpositive functions. *Journal of the American Statistical Association* 84: 710-6.

Thisted, R. A. 1988. *Elements of Statistical Computing*, New York: Chapman & Hall.

Tseng, G. C., Oh, M. K., Rohlin, L., Liao, J. and Wong, W. 2001. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Research* 29: 2549-57.

Tusher, V. G., R., T. and G., C. 2002. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences* 98: 5116-21.

Vaupel, J. W., Manton, K. G. and Stallard, E. 1979. The Impact of Heterogeneity in Individual Frailty on the Dynamics Mortality. *Demography* 16: 439-54.

Waagepetersen, R. and Sorensen, D. 2001. A tutorial on reversible jump MCMC with a view toward applications in QTL-mapping. *International Statistical Review* 69: 49-61.

Wagner, M. and Naik, D. N. 2007. Issues in mass spectrometry profiling in disease proteomics. In *Computational Methods in Biomedical Research*, eds. R. Khattree and D. N. Naik, this volume, pp. Wall, M. M. 2004. A close look at the spatial correlation structure implied by the CAR and SAR models. *Journal of Statistical Planning and Inference* 121: 311-24.

Wall, M. M. 2004. A close look at the spatial correlation structure implied by the CAR and SAR models. *Journal of Statsitical Planning and Inference* 121: 311-24.

Waller, L. A. and Carlin, B. P. and Xia, H. Gelfand, A. E. 1997. Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association* 92: 607-17.

West, M. 1985. Generalized linear models: scale parameters, outlier accommodation and prior distributions. In *Bayesian Statistics* 2: 531-57, Oxford: Oxford University Press.

West, M. and Harrison, P. J. and Migon, H. 1985. Dynamic generalised linear models and Bayesian forecasting (with discussion). *Journal of the American Statistical Association* 80: 73-97.

West, M. and Harrison, P. J. 1989. *Bayesian Forecasting and Dynamic Models*. New York: Springer-Verlag.

Wolfinger, R. D. and Kass, R. E. 2000. Nonconjugate Bayesian Analysis of Variance Component Models. *Biometrics* 56: 76874.

Wu, W., Black, M., Mumford, D., Gao, Y., Bienenstock, E., and Donoghue, J. 2003. A switching Kalman filter model for the motor cortical coding of hand motion. In *Proc. IEEE Engineering in Medicine and Biology Society Conference*, 2083-6.

Xue, X. and Brookmeyer, R. 1996. Bivariate frailty model for the analysis of multivariate survival time. *Lifetime Data Analysis* 2: 277-89.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J. and Speed, T. 2002. Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research* 30: e15.