Identifying Carriers of a Genetic Modifier Using Nonparametric Bayesian Methods

Peter D. Hoff
Richard B. Halberg
Alexandra Shedlovsky
William F. Dove
Michael A. Newton

ABSTRACT Animals in a certain population of mice each carry a mutant allele called $\star$ with probability one-half. This population is bred to a strain of mice which carry the *Min* allele of the *APC* gene, an allele which results in the development of intestinal tumors. Offspring from this cross are genotyped for *Min* and tumor counts are recorded. It is assumed that offspring carrying only *Min* have tumor counts distributed according to P_1, while offspring having both *Min* and the $\star$ allele have tumor counts distributed according to P_2, a probability measure assumed to be stochastically smaller than P_1. Presence of the *Min* allele is observable, but presence of the $\star$ allele is not. Given the tumor count data and assuming the stochastic ordering constraint, our goal is to estimate P_1, P_2 and the the unobserved genotype information. This is done by putting a nonparametric prior on the space of all pairs of stochastically ordered tumor count distributions, and computing posterior quantities of interest using MCMC.

1 Introduction

People with familial adenomatous polyposis (FAP) develop hundreds to thousands of benign tumors of the colon, which if untreated eventually progress to become carcinomas. The disease results from an inherited mutation in the adenomatous polyposis coli (*APC*) gene. The *Min* mutation in the mouse homologue of *APC* results in a phenotype very similar to human FAP. Mice with the *Min* mutation thus provide a model for studying this type of inherited colon cancer (Dietrich et al., 1993).

In a mutagenesis experiment, a mouse is obtained which shows signs of carrying a mutant allele at a modifier gene, suppressing the tumor-causing

effects of *Min*. In order to genetically map the location of the modifier gene, it is necessary to breed and identify a group of animals carrying the modifier allele. Although inheritance of the modifier is not directly observable, animals resulting from a breeding experiment carry the modifier with probabilities determined by the rules of Mendelian inheritance. Conditional upon the unobserved pattern of inheritance, each animal is modeled as having a tumor count sampled from either a carrier or a non-carrier probability distribution. Our goal is to estimate the two probability distributions and identify likely carriers and non-carriers of the modifier, assuming only that the tumor count distributions are stochastically ordered.

We take a nonparametric Bayesian approach, putting a prior on the space of pairs of stochastically ordered distributions. Such a prior can be constructed indirectly by putting a Dirichlet prior on the set of bivariate distributions of latent observations, members of this set having support only on ordered pairs of points. The marginals of such distributions will follow the stochastic ordering, and thus the Dirichlet prior on distributions of latent observations induces a prior on pairs of stochastically ordered distributions. This technique of modeling a collection of constrained distributions via an unconstrained latent distribution is discussed by Hoff (2000) in the context of maximum likelihood estimation.

Although construction of our prior is straightforward, computation of posterior quantities is quite difficult. We construct a Markov chain to generate approximate samples from the posterior. In order to achieve sufficient mixing in our sequence of posterior samples, our chain uses a combination of Gibbs and Hastings updates, based on full and partial conditioning (Besag, Green, Higdon and Mengersen, 1993).

2 Breeding Scheme

A kindred founder mouse is suspected of carrying an allele, referred to hereafter as $\star$, which suppresses the tumor-causing effects of *Min*. This kindred founder is bred to the *BTBR* strain of mice to produce a new population, members of which carry the $\star$ allele independently with probability one-half. Animals in this population are referred to as subkindred founders, for which presence or absence of $\star$ is unobserved.

Subkindred founders are bred to the *B6 Min/+* strain of mice, members of which carry one copy of the *Min* allele, causing intestinal tumor growth. From this cross, the resulting offspring carrying *Min* are identified by genotyping and their tumor counts are recorded. This group of mice is referred to as the NF population. Note that if a subkindred founder carries

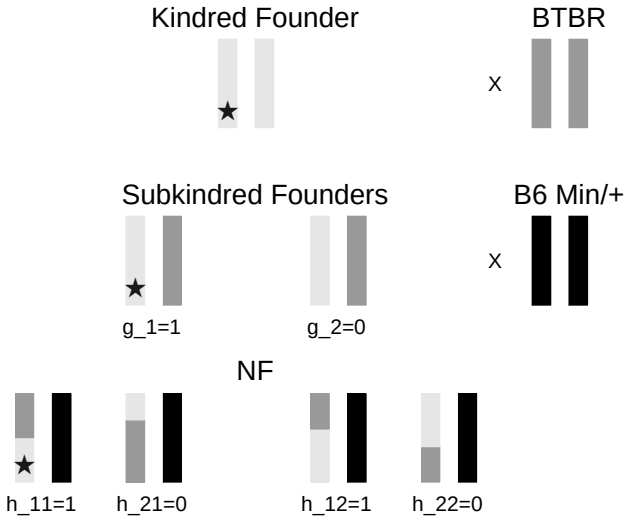

FIGURE 8.1. Basic Breeding Scheme

the $\star$ allele, then so will roughly half of its NF offspring.

This breeding scheme and model of inheritance are outlined in Figure 8.1. We let g_j be the random variable indicating presence or absence of the $\star$ allele in the jth subkindred founder, $j = 1, \ldots, m$, and note the Mendelian model of inheritance implies $g_1, \ldots, g_m$ are i.i.d. Bernoulli(1/2) random variables. One goal of this paper is to estimate the g_j's from the NF tumor count data. These estimates of carrier status will be used in future work to genetically map the location of the modifier gene.

Each NF animal inherits one set of chromosomes from its subkindred parent and one set from its *B6 Min/+* parent. Because of the possibility of chromosomal crossing-over, each chromosome inherited by an NF animal from its subkindred parent may be a mixture of chromosomal material from the kindred founder and the *BTBR* strain. In the region of the genome where $\star$ resides, the probability that a particular NF mouse has chromosomal material from the kindred founder is one-half. We denote the indicator of this event for the ith mouse in subkindred j as $h_{(i,j)}$. The $h_{(i,j)}$'s are i.i.d. Bernoulli(1/2) random variables, and are independent of the g_j's. The indicator of the event that mouse (i, j) has the $\star$ allele can be written as $x_{(i,j)} = g_j h_{(i,j)}$.

We note that some of the data analyzed in this paper were generated using slightly different breeding schemes. However, the basic structure of

the above model is applicable to all of them, and the carrier and non-carrier tumor count distributions should be common to all subkindred populations, regardless of the particular breeding scheme.

3 Latent Tumor Counts and Stochastic Ordering

Let P_1 and P_2 denote the tumor count distributions of NF animals which are non-carriers and carriers of $\star$ respectively, and let the set of all possible tumor counts be $\mathcal{Y}$. Our assumptions about $\star$ suggest a sample from P_1 is "probably larger" than a sample from P_2. One possible mathematical model for such an assumption is that P_1 is stochastically larger than P_2, that is $P_2(y,\infty) \leq P_1(y,\infty)\ \forall\, y$, in which case we write $P_2 \preceq P_1$.

Theorem 1: *$P_2 \preceq P_1$ if and only if there exists a measure P on $\mathcal{S} = \{s \in \mathcal{Y}^2 : s_2 \leq s_1\}$ such that P_1 and P_2 are the first and second marginals of P.*

The above result can be proven directly as by Lehmann (1986, Section 3.3), or can be seen as an application of a Choquet-type theorem, as described by Hoff (2000). Using this parametrization, an observation y distributed according to $P_k,\ k = 1, 2$ can be modeled as follows:

- Sample $s \sim P$;
- observe $y = s_k$.

We can think of s as being partially observed latent data, and y as the observed data. Estimating P_1 and P_2 subject to the stochastic ordering constraint can be done via unconstrained estimation of the measure P. In this way, a constrained estimation problem can be rewritten as an unconstrained missing-data problem, which is often easier to solve.

This parametrization provides a natural interpretation of the stochastic ordering constraint: We assume the tumor count y of each animal in our experiment is a deterministic function of $x \in \{0, 1\}$, the indicator of the presence of $\star$, and other unrecorded information $\omega \in \Omega$, so $y = y(\omega, x)$. If we assume the presence of $\star$ reduces tumor count, then it is natural to suppose $y(\omega, 1) \leq y(\omega, 0)\ \forall\, \omega$, i.e. all else being equal, the presence of $\star$ will not increase tumor count. Now define $s(\omega) = \{s_1(\omega), s_2(\omega)\} = \{y(\omega, 0), y(\omega, 1)\}$ as the vector of latent tumor counts. Any probability measure on Ω induces a canonical measure P on s so that $s_2 \leq s_1$ a.s. P. Furthermore, the marginals of P will satisfy the ordering $P_2 \preceq P_1$.

4 A Hierarchical Model

Our goal is to estimate P_1, P_2, and the missing subkindred genotype information $g_1, \ldots, g_m$ from the observed tumor count data. A nonparametric Bayesian approach involves a prior for (P_1, P_2) having support on all pairs of stochastically ordered measures on $\mathcal{Y}$. Such a prior is induced by the construction of a prior for the latent tumor count distribution P: If the support of the prior for P includes all possible distributions of ordered latent tumor counts in $\mathcal{Y}^2$, then by virtue of Theorem 1, the induced prior on the marginals has support on all pairs (P_1, P_2) such that $P_2 \preceq P_1$.

In this paper, our prior for P is based upon a simple parametric family of probability measures for latent tumor counts. Our uncertainty about the adequacy of the parametric family is quantified by assuming P is a sample from a Dirichlet process, centered around a base measure which is a member of the parametric family. A parametric prior on the base measure results in a nonparametric hierarchical prior for P.

4.1 A Parametric Model For Latent Tumor Counts

Suppose a set of cells in an organism have a certain probability of developing into tumors independently of one another. A model for total tumor count would then be a binomial distribution. Since the probability of tumorigenesis is typically quite small, and the number of cells in question is quite large, the binomial model of tumor counts can be well approximated by a Poisson model. Now suppose we are looking at a population of tumor counts, obtained from a population of organisms, each of whom have potentially different rates of tumorigenesis. Assuming a gamma prior for the population of rates, the resulting distribution of tumor counts follows a negative binomial model, a two parameter family of distributions with support on the nonnegative integers, with a density given by

$$p_{negbin}(s|\theta,\gamma) = \frac{\Gamma(s+\gamma)}{\Gamma(s+1)\Gamma(\gamma)} \left(\frac{\gamma}{\gamma+\theta}\right)^{\gamma} \left(\frac{\theta}{\gamma+\theta}\right)^{s}.$$

With this parametrization, $\mathrm{E}(s|\theta,\gamma) = \theta$, and $\mathrm{Var}(s|\theta,\gamma) = \theta(1+\theta/\gamma)$. Modeling tumor counts using the negative binomial distribution has been discussed before, for example in Drinkwater and Klotz (1981).

Our parametric model for latent tumor counts is as follows: We assume the tumor count s_1 of each non-carrier of $\star$ follows a negative binomial(θ,γ) distribution. We model tumor suppression by assuming each tumor that

would have developed without $\star$ develops with probability p in the presence of $\star$, independently of the other tumors. This implies the conditional distribution of the suppressed tumor count s_2 given s_1 is binomial(s_1, p), and so the resulting joint distribution of (s_1, s_2) has support on $s_2 \leq s_1$. It is interesting to note that, unconditional on s_1, s_2 is distributed according to a negative binomial$(p\theta, \gamma)$ distribution, and so p can be interpreted as the multiplicative effect of $\star$ on mean tumor count.

4.2 Nonparametric Extension

Our knowledge of tumorigenesis suggests the above model is reasonable, but we would like to relax the strict parametric assumptions. This can be done by using the Dirichlet prior: A Dirichlet prior $\mathcal{D}(\alpha P_0)$ is a probability measure on a space of distributions parametrized by a positive weight parameter α and a base measure P_0. A probability measure P sampled from $\mathcal{D}(\alpha P_0)$ is "centered" around P_0 in the sense that if $x_1, \ldots, x_n | P$ are i.i.d. observations from P, then marginally $\mathrm{E}[\mathrm{E}_P(f(x_i))] = \mathrm{E}_{P_0}[f(x_i)]$. However, such observations are marginally correlated, and in particular

$$\mathrm{E}\left(\sum (x_i - \bar{x})^2/(n-1)\right) = \sigma_0^2 \frac{\alpha}{\alpha + 1},$$

where σ_0^2 is the variance of a single observation under the base measure P_0. In fact, as $\alpha \to 0$, P converges to a point-mass measure with support on a random draw from P_0 (Sethuraman and Tiwari, 1981). See Ferguson (1973) or Blackwell and MacQueen (1973) for a more detailed account of the Dirichlet prior.

Our nonparametric model for the latent tumor count distribution P is a Dirichlet prior with a fixed α parameter and a random base measure. Our uncertainty about the base measure is quantified by a parametric prior π on a family of base measures, parametrized by

- the expectation of the non-suppressed tumor counts $\theta = E(s_1)$;
- the multiplicative effect p of $\star$, so that $p\theta = E(s_2)$;
- the expected sample variance of the non-suppressed tumor counts $\sigma^2 = E\left(\sum (s_{(i,j),1} - \bar{s}_1)^2/(n-1)\right)$.

More specifically, for a given value of $\phi = (\theta, p, \sigma^2)$, the base measure P_ϕ is given by

- $\gamma = \theta^2/(\sigma_0^2 - \theta)$, where $\sigma_0^2 = \sigma^2(\alpha + 1)/\alpha$;

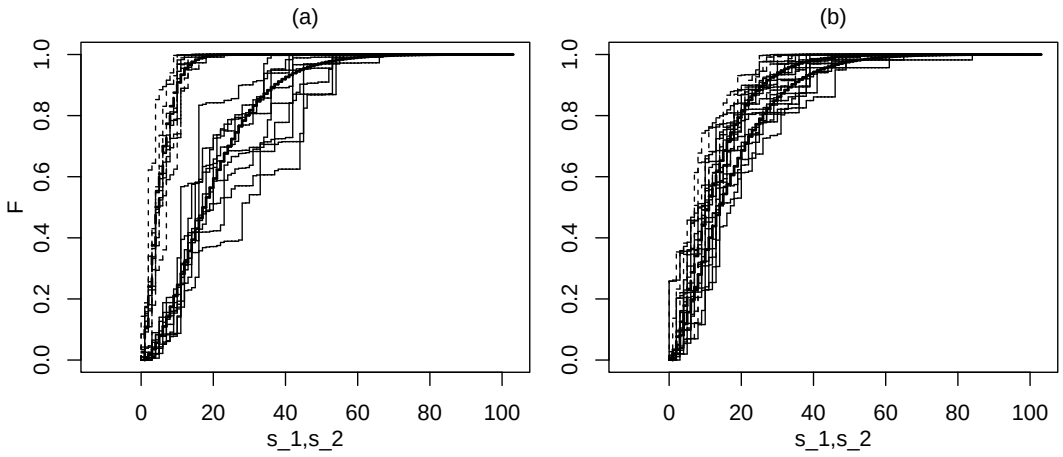

FIGURE 8.2. Marginal samples from $\mathcal{D}(\alpha P_\phi)$, with $\alpha = 10$ and (a) ϕ=(20,150,.25), (b) ϕ=(17,150,.75). Thick lines are marginals of P_ϕ, and solid and dashed lines represent the non-suppressed and suppressed groups respectively.

- $P_\phi(s_1) = p_{\text{negbin}}(s_1|\theta,\gamma)$;
- $P_\phi(s_2|s_1) = p_{\text{bin}}(s_2|s_1,p)$.

Given a prior π on ϕ, the complete hierarchical model is as follows:

- Hyperparameter: $\phi \sim \pi(\phi)$;
- Latent tumor count distribution: $P|\phi \sim \mathcal{D}(\alpha P_\phi)$;
- Latent observations: $s_{(1,1)},\ldots,s_{(n_m,m)}|P \sim$ i.i.d. P;
- Genotype Information: $h_{(1,1)},\ldots,h_{(n_m,m)},g_1,\ldots,g_m \sim$ i.i.d. Bernoulli(1/2);
- Observed data: $y_{(i,j)} = \begin{cases} s_{(i,j),1} & \text{if } h_{(i,j)}g_j = 0; \\ s_{(i,j),2} & \text{if } h_{(i,j)}g_j = 1. \end{cases}$

Animals with the *Min* allele without $\star$ have been well studied, populations of such mice having average tumor counts of roughly 20 and a population variance of about 150. We therefore use a gamma(40,0.5) prior for θ and gamma(150,1) prior for σ^2 to reflect our uncertainty about these parameters. The effect of $\star$ is not known; this uncertainty is quantified by a uniform prior for p on the interval $(0,1)$.

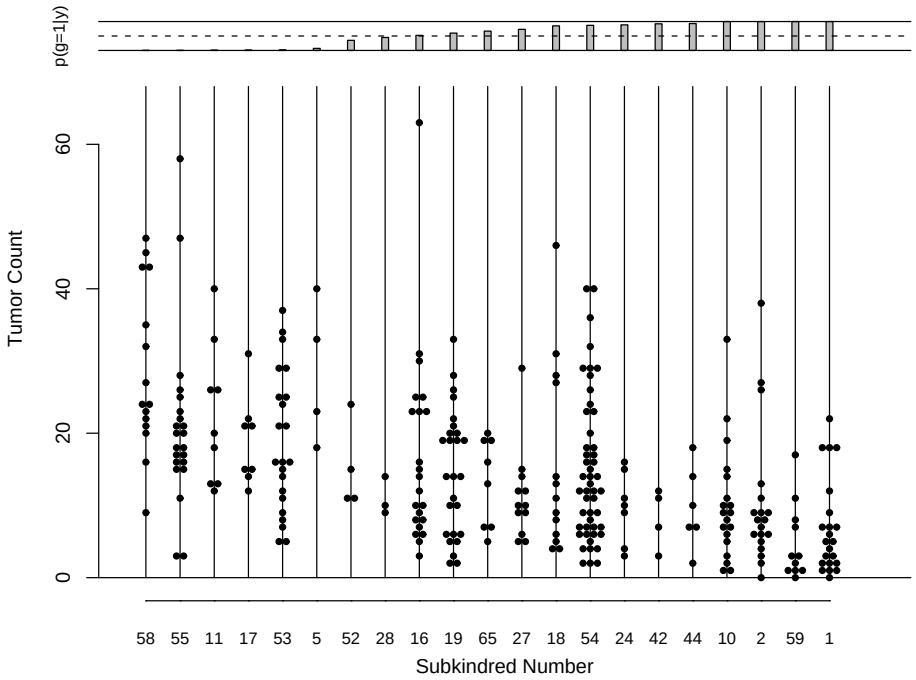

FIGURE 8.3. Some tumor count data.

The α parameter determines, among other things, how close the tumor count measure P is to the base measure P_ϕ for a given ϕ. For each of two different P_ϕ we have drawn 10 samples from a $\mathcal{D}(\alpha P_\phi)$ distribution with $\alpha = 10$ and plotted the resulting marginals in the two panels of Figure 8.2. Having studied other populations of such mice, we think this α-value of 10 roughly reflects our uncertainty about the fit of the parametric negative binomial/binomial model.

5 Data Analysis

Data were collected from 74 subkindreds, with tumor counts from 968 NF animals. Tumor counts ranged from zero to 79, with an average of 15.74 and a standard deviation of 11.51. Tumor count data from 21 subkindreds selected at random are shown in Figure 8.3. Each vertical line represents a subkindred, with dots plotted along a line representing the tumor counts of NF offspring from the corresponding subkindred founder.

Recall that a subkindred founder carrying $\star$ will pass the allele on to each of its NF offspring independently with probability one-half. We there-

fore expect tumor counts from such a subkindred to be an approximately equal mix of high and low values. Conversely, we expect mostly high tumor counts from animals in a subkindred lacking $\star$. With this in mind, we might categorize subkindred founders represented on the right-hand side of Figure 8.3 as likely carriers, and those on the left-hand side as likely non-carriers. One goal of this data analysis is to make our determination of carrier versus non-carrier status more precise.

5.1 Markov Chain Implementation

Given the observed tumor count data y, we wish to calculate posterior estimates of the tumor count distributions P_1 and P_2 (which are deterministic functions of P), and the subkindred genotype information $g = (g_1, \ldots, g_m)$. These posterior quantities of interest involve complicated integrals over high-dimensional spaces. Therefore, we approximate these integrals by empirical distributions of samples from a Markov chain whose stationary distribution is the desired posterior. To facilitate the sampling, we include the latent tumor counts $S = (s_{(1,1)}, \ldots, s_{(n_m,m)})$ and the parameter ϕ in the construction of our chain. Given current values (g^b, S^b, ϕ^b, P^b), one scan of the chain consists of

1. sampling $g^{b+1} \sim \pi(g|P^b, \phi^b, y) = \pi(g|P^b, y)$, a distribution of independent Bernoulli random variables;
2. sampling $S^{b+1} \sim \pi(S|g^{b+1}, P^b, \phi^b, y) = \prod_{i,j} \pi(s_{(i,j)}|g_j^{b+1}, P^b, y_{(i,j)})$, in which $s_{(i,j)}^{b+1}$ is distributed as P^b conditional on $s_{(i,j),1}^{b+1} = y_{(i,j)}$ if $g_j^{b+1} = 0$, and is distributed as P^b conditional on at least one component of $s_{(i,j)}^{b+1}$ being equal to $y_{(i,j)}$ if $g_j^{b+1} = 1$;
3. sampling ϕ^* from a symmetric random walk distribution, and accepting ϕ^* as ϕ^{b+1} with probability $\frac{\pi(\phi^*|g^{b+1},S^{b+1},y)}{\pi(\phi^b|g^{b+1},S^{b+1},y)} = \frac{\pi(S^{b+1}|\phi^*)\pi(\phi^*)}{\pi(S^{b+1}|\phi^b)\pi(\phi^b)}$;
4. sampling $P^{b+1} \sim \pi(P|S^{b+1}, g^{b+1}, \phi^{b+1}, y) = \mathcal{D}(\alpha P_{\phi^{b+1}} + n\hat{P}_{S^{b+1}})$, a Dirichlet distribution where $\hat{P}_{S^{b+1}}$ is the empirical distribution of the current state of S.

The updates (1) and (3) for g and ϕ are based on partial conditionals, that is, the conditional distributions given some, but not all, of the current values of the other components. Such partial conditioning is justified by noting that the above sampling scheme is equivalent to

Step 1: a Gibbs update for (g, S);

Step 2: a Gibbs update for S;

Step 3: a Hastings update for (ϕ, P);

Step 4: a Gibbs update for P.

The equivalence can be seen via the following general argument: Suppose we wish to estimate a generic joint distribution $\pi(x, y, z)$ by MCMC methods. In some cases, it may be more desirable to update x based on $\pi(x|z)$ rather than the full conditional $\pi(x|y, z)$. This can be done by sampling an x^* from a desired proposal distribution $J_1(x^*|x, z)$, then "pretending" to sample y^* from $\pi(y|x^*, z)$. The proposal (x^*, y^*) is then accepted with probability

$$\begin{aligned}\frac{\pi(x^*, y^*|z)}{\pi(x, y|z)}\frac{J(x, y|x^*, y^*, z)}{J(x^*, y^*|x, y, z)} &= \frac{\pi(x^*|z)\pi(y^*|x^*, z)}{\pi(x|z)\pi(y|x, z)}\frac{\pi(y|x, z)J_1(x|x^*, z)}{\pi(y^*|x^*, z)J_1(x^*|x, z)} \\ &= \frac{\pi(x^*|z)J_1(x|x^*, z)}{\pi(x|z)J_1(x^*|x, z)}.\end{aligned}$$

By using a full conditional for y^* and a proposal distribution for x^* that doesn't depend on y (for example a partial conditional as in Step 1 above, or a symmetric random walk as in Step 3), we have ensured our acceptance probability of (x^*, y^*) doesn't depend on y^*. Therefore, y^* doesn't actually need to be generated at this stage; instead, it can be updated at the next stage, using a potentially different proposal mechanism. For a more detailed discussion of partial conditioning, see Besag, Green, Higdon, and Mengersen (1995, Appendix 2). We base our Markov chain on partial conditionals for the reasons given below.

To improve mixing: Although the chain based on the full conditionals is irreducible, it doesn't mix very well. This is because the conditional distribution of g given S, P and y is often degenerate: Consider a single subkindred j whose founder has unknown genotype g_j. Note that $g_j = 0$ implies the event $E_j = \{s_{(i,j),1} = y_{(i,j)}, i = 1, \ldots, n_j\}$, i.e. if the subkindred founder does not carry $\star$, then the tumor counts of its offspring are non-suppressed. On the other hand, $E_j^c \Rightarrow \{g_j = 1\}$, so the full conditional of g_j is a point mass at one if E_j does not hold. Given $g_j^b = 1$, sampling an S^b to satisfy E_j is possible but extremely unlikely. This in turn makes the probability $\Pr(\mathrm{g_j^{b+1}} = 0|\mathrm{g_j^b} = 1)$ very small, and leads to poor mixing. This difficulty is avoided by sampling g^{b+1} conditional on P^b and y only.

To reduce calculations and numerical errors: In the case of the Hastings step for ϕ, one would typically base the acceptance probability of ϕ^* on $\pi(\phi^*|P, S, g, y)$, which reduces to $\pi(\phi^*|P)$. Since $P|\phi^*$ is distributed as a Dirichlet process, computing this acceptance probability involves calculating $\Gamma(\alpha P_{\phi^*}(s))$ for all possible latent tumor counts s. Many values of $P_{\phi^*}(s)$ will be extremely small, making the calculation of the Gamma function prone to numerical errors. On the other hand, the conditional distribution of ϕ^* given only S involves computing $\Gamma(\alpha P_{\phi^*}(s))$ only for those values of s occurring in the sampled set of the latent tumor counts S. As $P_{\phi^*}(s)$ is typically larger for sampled values of s, this modification of the sampling scheme tends to reduce not only the number, but also the magnitude of the errors made in computing the Gamma function.

Finally, we note the updates for ϕ can be done component by component. That is, proposals and acceptances can be made separately for θ,γ, and p, and so after one scan of the chain, ϕ^{b+1} could be the same as ϕ^b, or could differ at one, two, or three component values. This component by component method of updating was used to make the inference in the following section.

5.2 Posterior Inference

For the purpose of data analysis, all tumor count distributions were conditioned to lie on the integers from zero to ninety. The sampling scheme described above was coded in the C++ programming language, and was used to generate four chains of 200,000 scans each, recording output every 100th scan. The starting values of P^0 for the four chains were generated by sampling P^0 from $\mathcal{D}(\alpha P_{\phi^0})$, using four different values of ϕ^0, given in Table 8.1.

The output of the chain is very high-dimensional: For diagnostics we only report on sequences of mean tumor count for the non-suppressed and suppressed groups, $E(s_1|P)$ and $E(s_2|P)$. As can be seen in Figure 8.4, after about 20,000 scans the four separate sequences of $E(s_1|P)$ and $E(s_2|P)$ seem to have converged to similar distributions. We delete the first 50,000 scans from each chain to allow for burn in, and compute the sample acf values from the remaining 4×1500 scans (150,000 scans subsampled every 100th scan), given in Table 8.2.

The 4×1500 scans recorded after burn in were used to compute posterior quantities of interest, some of which are displayed in Figure 8.5. The first panel shows the posterior mean CDF's of the two stochastically ordered groups in heavy lines, with confidence bands in lighter lines. The

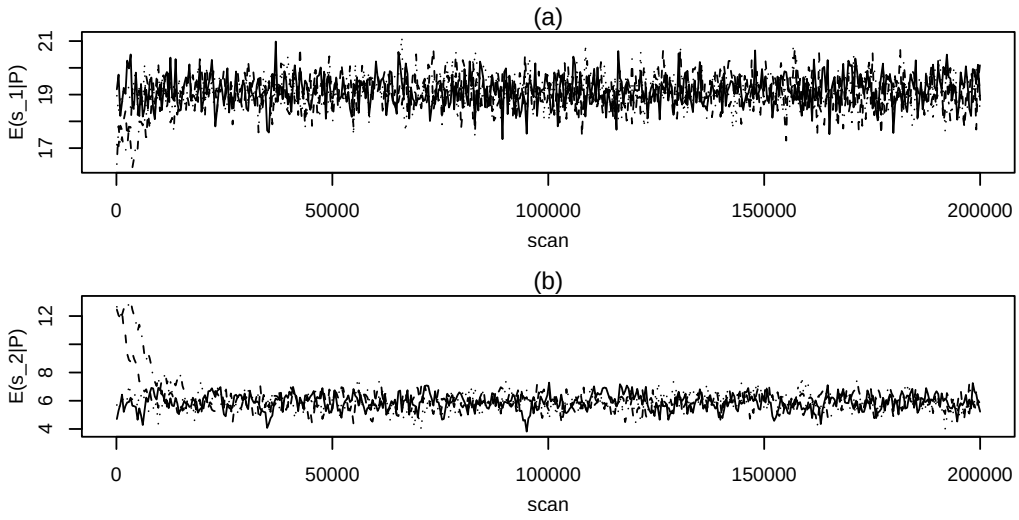

FIGURE 8.4. Sequences of (a) $E(s_1|P)$ and (b) $E(s_2|P)$ for four different chains.

Chain #	θ	σ^2	p
1	20	125	0.25
2	20	175	0.25
3	17	125	0.75
4	17	175	0.75

TABLE 8.1. Starting Values of $\phi^0 = (\theta, \sigma^2, p)^0$.

confidence bands represent the range of the CDF's saved from the chain, that is every 100th sample after the first 50,000 scans. The second panel gives a contour plot of the joint posterior density of $E(s_1|P)$ and $E(s_2|P)$. The contour lines represent highest posterior density regions of 20, 50, 80, 90, and 95 percent probability. The posterior means of these parameters are 19.11 and 5.89 respectively, with posterior standard deviations of 0.34 and 0.33 (based on weighted averages of within-chain and between-chain variances). The standard deviations of the tumor count distributions are estimated as 11.37 and 4.15. The third panel gives the marginal posterior distribution of the multiplicative effect p of $\star$ in the base model, which has a posterior mean and mode of .31, and a posterior standard deviation of .034. These three plots show the estimated effect of $\star$ to be quite large, giving about a 70% reduction in mean tumor count between the two populations. This is an important result, as an allele with such a large effect is

	Lag 100	Lag 500	Lag 1000	Lag 5000
acf($E(s_1\|P)$)	0.0590	0.0149	-0.0277	-0.0107
acf($E(s_2\|P)$)	0.7549	0.4637	0.2524	0.0102

TABLE 8.2. Sample Autocorrelation

of biological significance and warrants further study.

Another important piece of output from the Markov chain is the posterior distribution of the subkindred genotypes $g_1, \ldots, g_m$ (the posterior expectation of these variables for some subkindreds are plotted along the top of Figure 3). This output allows us to identify likely carriers and non-carriers of $\star$, which in turn will aid us in the next stage of inquiry—mapping the location of the modifier gene in the mouse genome.

6 Discussion

We have developed a model which allows us to measure the effect of a modifier allele on intestinal tumor count. From our analysis, we estimate the effect of the $\star$ allele to be a 70% reduction in the number of intestinal tumors in *Min* mice. Our model also identifies likely carriers and non-carriers of $\star$, which will allow us to map the location of the modifier gene once genetic marker data has been gathered.

Our model for suppressed and non-suppressed tumor count distributions involves a nonparametric prior on the space of pairs of stochastically ordered distributions. The prior is based on a Dirichlet distribution centered around a parametric negative binomial/binomial model for latent tumor counts, which allows for model flexibility while retaining a degree of smoothness. We have selected a fixed value of the weight parameter α in our Dirichlet prior, claiming that $\alpha = 10$ represents our prior beliefs about the tumor count sampling distributions. These prior beliefs are based on knowledge of tumor count distributions of similar mouse populations. Nevertheless, the analysis in this paper was redone with fixed α-values of 1 and 100, giving results very similar to those presented in this paper. Alternatively, we could have put a prior on α (as discussed in Escobar and West (1995)), although the insensitivity of the results to the choice of fixed α suggests this is unnecessary.

The model discussed in this paper can be extended to data analyses where there is an intuitive partial ordering on more than two subpopulations of the dataset. Modeling such an ordering can be accomplished via a multivariate distribution P for latent observations s, such that the com-

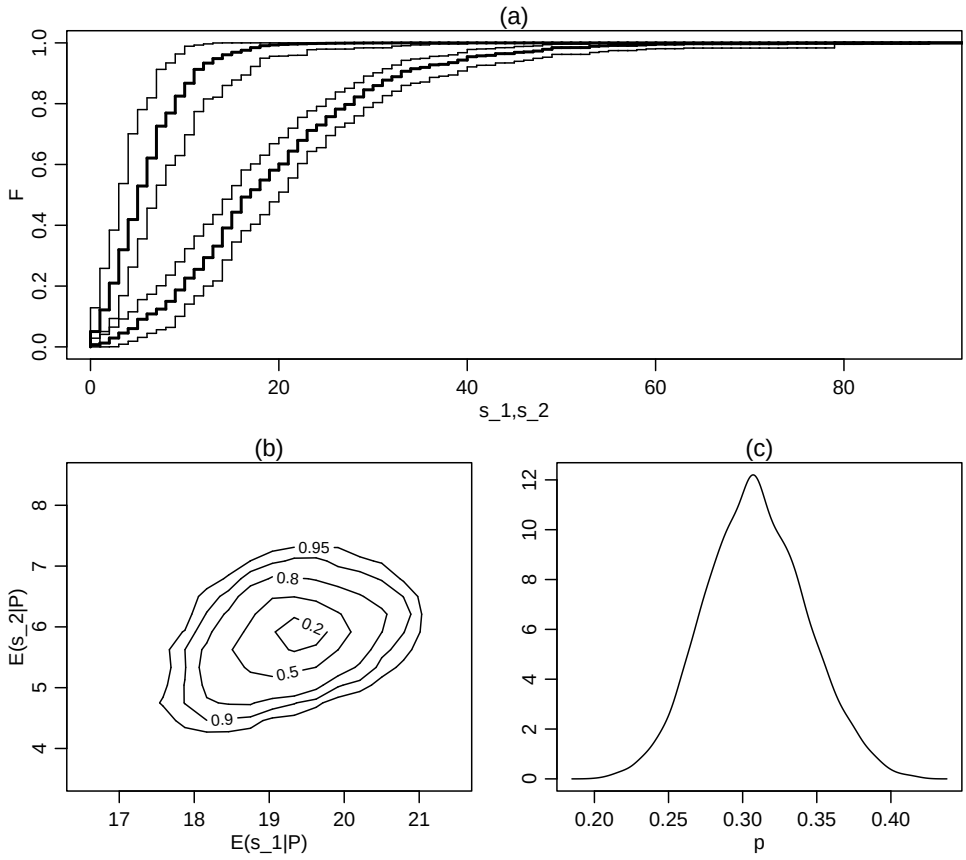

FIGURE 8.5. Posterior Quantities: (a) Bayes estimates of tumor count CDF's; (b) Posterior joint distribution of $E(s_1|P), E(s_2|P)$; (c) Posterior distribution of p.

ponents of s are ordered a.s. P according to the desired partial ordering. Such a model is presented in Hoff (2000) for a collection of four partially ordered distributions.

Acknowledgments

Thanks to Linda Clipson of the McArdle Laboratory of Cancer Research, University of Wisconsin-Madison, for her help in maintaining these data. This research was supported in part by National Cancer Institute grants T32-CA09565-09 for the first author, F32-CA77946 for the second author, R37-CA63677 for the third and fourth authors, and R29-CA64364-01 for the fifth author.

References

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995). Bayesian Computation and Stochastic Systems (with discussion). *Statistical Science* **10** , 3-66.

Blackwell, D., and MacQueen, J.B. (1973). Ferguson Distributions via Pólya Urn Schemes. *The Annals of Statistics* **1**, 353-355.

Dietrich, W.F., Lander, E.S., Smith, J.S., Moser, A.R., Gould, K.A., Luongo, C., Borenstein, N., and Dove, W. (1993). Genetic Identification of *Mom-1*, a Major Modifier Locus Affecting *Min*-induced Intestinal Neoplasia in the Mouse. *Cell* **75**, 631-639.

Drinkwater, N.R., and Klotz, J.H. (1981). Statistical Methods for the Analysis of Tumor Multiplicity Data. *Cancer Research* **41**, 113-119.

Escobar, M.D., and West, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association* **90**, 577-588.

Ferguson, T.S. (1973). A Bayesian Analysis of some Nonparametric Problems. *The Annals of Statistics* **1**, 209-320.

Hoff, P.D. (2000). Constrained Nonparametric Maximum Likelihood via Mixtures. *The Journal of Computational and Graphical Statistics*, To appear.

Lehmann, E.L. (1986). *Testing Statistical Hypotheses*. Singapore: John Wiley and Sons.

Sethuraman, J., and Tiwari, R.C. (1982). Convergence of Dirichlet Measures and the Interpretation of Their Parameter. In *Statistical Decision Theory and Related Topics III*, eds. S.S. Gupta and J.O. Berger. New York: Academic Press, 305-315.