RC23521 (W0502-054) February 8, 2005
Mathematics

IBM Research Report

NHPP Models for Categorized Software Defects

Zhaohui Liu, Nalini Ravishanker
Department of Statistics
University of Connecticut
Storrs, CT 06269

Bonnie K. Ray
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

NHPP Models for Categorized Software Defects

Zhaohui Liu, Nalini Ravishanker
Department of Statistics
University of Connecticut
Storrs CT 06269
zhl97002@yahoo.com, nalini@stat.uconn.edu

Bonnie K. Ray
Mathematical Sciences Department
IBM Watson Research Center
P.O. Box 218, Yorktown Heights, NY 10598
bonnier@us.ibm.com

January 11, 2005

NHPP Models for Categorized Software Defects

Key Words: Bayesian inference, conditional multinomial, latent variables, software engineering

Abstract: We develop NHPP models to characterize categorized event data, with application to modeling the discovery process for categorized software defects. Conditioning on the total number of defects, multivariate models are proposed for modeling the defects by type. A latent vector autoregressive structure is used to characterize dependencies among the different types. We show how Bayesian inference can be achieved via MCMC procedures, with a posterior prediction-based L-measure used for model selection. The results are illustrated for defects of different types found during the System Test phase of a large operating system software development project.

1. Introduction

Nonhomogeneous Poisson process (NHPP) models are commonly used to characterize event data collected over time when the expected number of events per time unit is non-constant. In particular, a number of different NHPP models have been developed for characterizing software reliability based on the observed number of defects discovered over time. Some of these include the models of Jelinski and Moranda [1], Goel and Okumoto [2] and Yamada, Ohba, and Osaki [3]. In the context of software reliability, the NHPP growth curve model of Kuo and Yang (see [4]) assumes that there is a Poisson distributed random number of defects N remaining in the software at the start of the model time frame ($t = 0$), that no new defects are introduced at each repair, and that the n failure times observed up to time t can be taken to be the first n order statistics from N independent and identically distributed ($i.i.d$) observations having probability distribution $f(t)$. Singpurwalla and Wilson [5] provide a comprehensive reference for statistical techniques useful for software reliability.

Almost all of the published literature on software reliability modeling has focused on characterizing a product's reliability based on all defects considered together, not distinguished by type. However, discovered defects are usually classified into different categories, such as severity, impact, or trigger event. Since different types of defects are expected to have different impacts on the ultimate software reliability, it is enlightening to model defects as classified into different types, in addition to the model that characterizes the overall defect discovery process. For instance, the Orthogonal Defect Classification (ODC) scheme was developed at IBM to characterize different types of defects found during the software development process. A few papers have discussed the relationship between type of defect and reliability growth, and have developed individual growth models to understand the evolution of defect types over time (see Chillarege *et al.* [6] and Chillarege and Biryani [7]). Ray, Bhandari and Chillarege [8] discussed a reliability growth model that can explicitly incorporate relationships for two defect types, the rationale being that certain defects cannot be discovered until other defects of a different type are first found. Ray, Liu and Ravishanker [9] described a software reliability characterization for typed defects using a dynamic growth curve formulation, which allows model parameters to vary as a

function of covariate information. In the case of categorized defects, the covariates used were the volumes of defects found in previous time periods of a different type.

In this paper we propose a different approach. Instead of considering the types interacting with each other simultaneously, we model the evolution of the proportion of different types of defects through a latent variable model with autoregressive structure to characterize dynamic interactions among the types. Although we motivate the model using categorized defects in the context of software reliability, this type of model may also be of interest in other applications. In survival analysis, for example, patients are routinely classified into different groups, with interest centered on the number of events experienced over time by each group. In a business context, it may be of interest to model the number of events of a particular type observed by a company, where some dependencies exist between the proportions at a given time. An alternate framework using a random coefficient autoregressive model was discussed in Singpurwalla and Soyer[10].

The remainder of the paper is organized as follows. After a brief background discussion, we introduce in Section 2 a bivariate NHPP model and its multivariate extension. In Section 3, we present Bayesian modeling details that enable full statistical inference. As illustration, we present estimation results for simulated data in Section 4, with application to actual software development defect data in Section 5. Section 6 concludes.

2. Conditional Multinomial NHPP Models

Let Y_{t_j} represent the number of events observed in the interval $[t_{j-1}, t_j)$ for $j = 1, \cdots, T$. The time intervals $[t_{j-1}, t_j), j = 1, \cdots, T$ typically represent equally spaced days, weeks, months, *etc.*, although they can, in general, be of varying lengths. The mean number of events m_{t_j} in $[t_{j-1}, t_j)$ is given by

$$m_{t_j} = \theta(F(t_j) - F(t_{j-1})), \tag{1}$$

where θ is the expected total number of events, assumed to be finite, and $F(t_j)$ represents the cumulative distribution function of the assumed time to event distribution.

A flexible distribution which can accommodate increasing or decreasing hazard rates is the Weibull distribution

$$f(t) = \alpha\beta t^{\alpha-1} \exp(-\beta t^{\alpha}), \quad \alpha > 0, \beta > 0, t > 0, \tag{2}$$

which has decreasing, constant, or increasing hazard rate for $\alpha < 1, = 1, > 1$, respectively; β can be interpreted as the *defect discovery rate*, while α has been interpreted in relation to the customer usage rate Kenney [11] The Weibull model reduces to an exponential model when $\alpha = 1$. Under a Weibull assumption, Equation (1) has form

$$m_{t_j} = \theta(F(t_j) - F(t_{j-1})) = \theta[\exp(-\beta t_{j-1}^{\alpha}) - \exp(-\beta t_j^{\alpha})]. \tag{3}$$

Suppose that these events are categorized into K types, so that $Y_{t_j}^{(k)}$ denotes the observed number of events of type k, $k = 1, \cdots, K$. We assume that the probability for each of the

K types of events, conditional on the total number of events within the time interval $[t_{j-1}, t_j)$ follows a multinomial distribution. That is, let $p_{t_j}^{(k)}$ represent the probability of a Type k event in $[t_{j-1}, t_j)$, $k = 1, \cdots, K$. Given that $Y_{t_j}^{(k)}$ is the number of events of Type k and Y_{t_j} is the total number of all types of $[t_{j-1}, t_j)$, it follows from well-known probability results (see Ross [12]) that the joint probability of the observed number of categorized events in $[t_{j-1}, t_j)$ is

$$P(Y_{t_j}^{(k)} = y_{t_j}^{(k)},\ k = 1, \cdots,\ K) = P(Y_{t_j}^{(k)} = y_{t_j}^{(k)} \mid Y_{t_j} = \sum_{k=1}^{K} y_{t_j}^{(k)}\ ) P(Y_{t_j} = \sum_{k=1}^{K} y_{t_j}^{(k)}\ ) \quad (4)$$

$$= \frac{(\sum_{k=1}^{K} y_{t_j}^{(k)})!}{\prod_{k=1}^{K} y_{t_j}^{(k)}!} \prod_{k=1}^{K} p_{t_j}^{y_{t_j}^{(k)}} \frac{e^{-m_{t_j}} m_{t_j}^{\sum_{k=1}^{K} y_{t_j}^{(k)}}}{(\sum_{k=1}^{K} y_{t_j}^{(k)})!} \quad (5)$$

where m_{t_j} was given in (1).

The likelihood function for the model parameters is obtained as the product of the joint probabilities in (5) over all time intervals. Let $\mathbf{Y} = \{Y_{t_j}, j = 1, \cdots, T\}$ denote the total number of observed events up to time T, and let $Y^{(k)}$ denote the number of events of Type $k, k = 1, \cdots, K$. Let $\mathbf{p}_{t_j} = (p_{t_j}^{(1)}, \cdots, p_{t_j}^{(K)})$, $\mathbf{p} = \{p_{t_j}^{(k)}, k = 1, \cdots, K; j = 1, \cdots, T\}$ and suppose Ψ denotes all the model parameters. The likelihood function for the observed data is

$$\mathrm{L}(\mathbf{Y}; \Psi) = \prod_{j=1}^{T} \frac{m_{t_j}^{Y_{t_j}} \exp(-m_{t_j})}{Y_{t_j}!}, \quad (6)$$

so that the logarithm of the observed-data likelihood is

$$\log\ f(\theta, \beta, \alpha, \mathbf{p}; \mathbf{Y}) \propto -\sum_{j=1}^{T} m_{t_j} + \sum_{j=1}^{T} Y_{t_j}\ \log\ m_{t_j} + \sum_{j=1}^{T} \sum_{k=1}^{K} Y_{t_j}^{(k)}\ \log\ p_{t_j}^{(k)}, \quad (7)$$

where $\mathbf{p}$ can be a function of time.

We discuss two multivariate model formulations for categorized events. The models described in Section 2.1 assume that the probability of a Type k event is constant over time, and further assumes no interaction between the probability of a Type k event and a Type l event for $k \neq l$, $k, l = 1, \cdots, K$. However, this assumption is restrictive, since it is reasonable to expect that categorized events are correlated both across categories and over time. For instance, certain defect types might trigger, or at least considerably increase, the possibility of other types. In Section 2.2 we describe models that explicitly capture such dependence via a vector autoregressive model for the logit transformations of the probabilities of categorized events.

2.1. Constant Proportion Models

We assume that $p_{t_j}^{(k)} = p^{(k)}$ in $[t_{j-1}, t_j)$, for $j = 1, \cdots, T$. Then, the observed-data likelihood in (7) simplifies as

$$\log\, f(\theta, \beta, \alpha, \mathbf{p}; \mathbf{Y}) \propto -\sum_{j=1}^{T} m_{t_j} + \sum_{j=1}^{T} Y_{t_j} \log\, m_{t_j} + \sum_{k=1}^{K} Y^{(k)} \log\, p^{(k)} \tag{8}$$

Since $\sum_{k=1}^{K} p^{(k)} = 1$, we only need to model the first $K-1$ probabilities. In addition to three parameters for the Weibull model for the total number of events, viz., θ, β, α, we have $K-1$ probabilities, $p^{(1)}, \cdots, p^{(K-1)}$, so that the vector of model parameters is $\Psi = (\theta, \beta, \alpha, p^{(1)}, \cdots, p^{(K-1)})$. We refer to this as Model MV1.

In many applications, there are only two types of events, for which the model simplifies. Let $p_{t_j}^{(1)}$ and $1 - p_{t_j}^{(1)}$ respectively denote the probability of Type 1 events and Type 2 events in $[t_{j-1}, t_j)$. Assume that $p_{t_j}^{(1)}$ remains constant for Type 1 events across all time intervals, i.e., $p_{t_j}^{(1)} = p^{(1)}$, for $j = 1, \cdots, T$; for simplicity of notation, we have dropped the superscript for $p^{(1)}$. Assuming that the total number of events follows a Weibull NHPP, the observed-data log likelihood function is

$$\log f(\theta, \beta, \alpha, p; \mathbf{Y}) \propto -\sum_{j=1}^{T} m_{t_j} + \sum_{j=1}^{T} Y_{t_j} \log\, m_{t_j} + Y^{(1)} \log\, p + Y^{(2)} \log\, (1-p) \tag{9}$$

where $\Psi = (\theta, \beta, \alpha, p)$. We refer to this as Model BV1.

2.2. Stochastic Proportion Models

Let

$$\begin{aligned} p_{t_j}^{(k)} &= \frac{\exp(q_{t_j}^{(k)})}{1 + \sum_{k=1}^{K-1} \exp(q_{t_j}^{(k)})},\ k = 1, \cdots, K-1 \\ p_{t_j}^{(K)} &= \frac{1}{1 + \sum_{k=1}^{K-1} \exp(q_{t_j}^{(k)})} \end{aligned} \tag{10}$$

denote the multinomial logit transformation for $\mathbf{p}_{t_j}$, where

$$q_{t_j}^{(k)} = \text{logit}(p_{t_j}^{(k)})$$

is the logit function. Assume that $\mathbf{q_{t_j}} = (q_{t_j}^{(1)}, \cdots, q_{t_j}^{(K-1)})'$ follows a stationary VAR(1) process, i.e.,

$$\mathbf{q}_{t_j} = \Phi_0 + \Phi_1\ \mathbf{q}_{t_{j-1}} + \epsilon_{t_j} \tag{11}$$

where Φ_0 is a $(K-1)\times 1$ vector, Φ_1 is a $(K-1)\times(K-1)$ matrix, and ϵ_{t_j} iid $\sim \mathrm{N}_{K-1}(0,\Sigma_\epsilon)$.

Suppose $\mathbf{Q}=(\mathbf{q}_1',\cdots,\mathbf{q}_T')'$, $\mathbf{x}_t'=(1,\mathbf{q}_{T-1}')$, $\mathbf{X}=(\mathbf{x}_1',\cdots,\mathbf{x}_T')'$, $\epsilon=(\epsilon_1',,\cdots,\epsilon_T')$, and $\Phi=(\Phi_0',\Phi_1)'$. The latent component can be compactly written as

$$\mathbf{Q}=\mathbf{X}\Phi+\epsilon$$

Using standard results in multivariate statistics (see Sun and Ni [13])

$$\log\, f(\mathbf{q}|\Phi,\Sigma_\epsilon)\propto -\log\,(|\Sigma_\epsilon|^{\frac{T}{2}})-\tfrac{1}{2}\textstyle\sum(\mathbf{q}_{t_j}'-\mathbf{x}_{t_j}'\Phi)\Sigma_\epsilon{}^{-1}(\mathbf{q}_{t_j}'-\mathbf{x}_{t_j}'\Phi)'.$$

Since $f(\mathbf{Y},\mathbf{Q}|\Delta_\mathbf{1},\Delta_\mathbf{2})=f(\mathbf{Y}|\Delta_\mathbf{1})\cdot f(\mathbf{Q}|\Delta_\mathbf{2})$ for arbitrary parameters Δ_1 and Δ_2, the complete data likelihood is

$$\begin{aligned}\log f(\theta,\beta,\alpha,\Phi,\Sigma_\epsilon;\mathbf{Y},\mathbf{Q})\propto &-\sum_{j=1}^{T}m_{t_j}+\sum_{j=1}^{T}Y_{t_j}\log\, m_{t_j}-\sum_{j=1}^{T}Y_{t_j}\log\,(1+\sum_{k=1}^{K-1}\exp(q_{t_j}^{(k)}))+\\ &\sum_{j=1}^{T}\sum_{k=1}^{K-1}y_{t_j}^{(k)}q_{t_j}^{(k)}-\log\,(|\Sigma_\epsilon|^{\frac{T}{2}})-\frac{1}{2}\sum(\mathbf{q}_{t_j}{}'-\mathbf{x}_{t_j}'\Phi)\Sigma_\epsilon{}^{-1}(\mathbf{q}_{t_j}'-\mathbf{x}_{t_j}'\Phi)'\end{aligned}\tag{12}$$

Here, $\Psi=(\theta,\beta,\alpha,\Phi,\Sigma_\epsilon)$. We refer to this as Model MV2.

If there are only two types of events, the latent variable model is

$$q_{t_j}=\phi_0+\phi_1 q_{t_{j-1}}+u_{t_j}\tag{13}$$

where $u_{t_j}\sim$ iid $\mathrm{N}(0,\sigma^2)$ and

$$q_{t_j}=\mathrm{logit}(p_{t_j})=\log(\tfrac{p_{t_j}}{1-p_{t_j}}),\ j=1,\cdots,T$$

is the standard logit transformation for p_{t_j}.

Here, we have assumed that the q_{t_j}'s follow a stationary AR(1) model. Inclusion of an intercept term ϕ_0 in (13) implies allowing for p_{t_j}'s to be different from 0.5. The AR(1) model is chosen for its ability to capture simple dependencies both across categories and within a category over time, but of course more complex dependencies could be modeled using higher order autoregressive or moving-average models.

The complete-data likelihood is

$$\log\, f(\mathbf{Y},\mathbf{q}|\theta,\beta,\alpha,\phi_0,\phi_1,\sigma^2)=\log f(\mathbf{Y}|\theta,\beta,\alpha)+\log f(\mathbf{q}|\phi_0,\phi_1,\sigma^2),$$

so that

$$\begin{aligned}\log f(\mathbf{Y},\mathbf{q}|\theta,\beta,\alpha,\phi_0,\phi_1,\sigma^2)\propto &-\sum_{j=1}^{T}m_{t_j}+\sum_{j=1}^{T}Y_{t_j}\log m_{t_j}+\sum_{j=1}^{T}y_{t_j}^{(1)}q_{t_j}\\ -&\sum_{j=1}^{T}Y_{t_j}\log(1+\exp(q_{t_j}))-T\log\sigma+\frac{1}{2}\log(1-\phi_1^2)\\ -&\frac{1}{2\sigma^2}\{(1-\phi_1^2)(q_{t_1}-\frac{\phi_0}{1-\phi_1})^2+\sum_{j=1}^{T-1}(q_{t_{j+1}}-\phi_0-\phi_1 q_{t_j})^2\}\end{aligned}\tag{14}$$

Here, $\Psi = (\theta, \beta, \alpha, \phi_0, \phi_1, \sigma^2)$. We refer to this as Model BV2.

Although maximum likelihood estimates of model parameters may be obtained via numerical optimization, a Bayesian approach has the advantage of incorporating prior information about model parameters, making inference on the latent variables, and providing reliable estimates with small samples. We detail the Bayesian inference scheme for the proposed models in the next section.

3. Bayesian inference

Bayesian inference is based on the posterior distribution of the model parameters, which is proportional to the product of the likelihood function and the prior, as shown below for each model. When the posterior distribution is not analytically tractable, Markov chain Monte Carlo algorithms facilitate inference via simulated samples from the complete conditional distributions of the parameters (see Gelfand and Smith [14] or Robert and Casella [15] for more details). For brevity, we will use $f(\Lambda|\cdots)$ to generically denote the complete conditional distribution of Λ, given all other model parameters and the data.

The Weibull NHPP model parameters are reparametrized by letting $c = \log(\theta)$, $d = \log(\beta)$ and $e = \log(\alpha)$. Under each of the models discussed in section 2, we assume normal priors for these parameters, i.e., $N(\mu_c, \sigma_c^2)$ for c, $N(\mu_d, \sigma_c^2)$ for d, and $N(\mu_e, \sigma_e^2)$ for e. Under Model MV1, we assume Jeffrey's prior for $\mathbf{p}$, which is the Dirichlet distribution $D_K(\frac{1}{2}, \cdots, \frac{1}{2})$. The corresponding prior under Model BV1 is the Beta$(\frac{1}{2}, \frac{1}{2})$ distribution. Under Model MV2, the priors for Φ and Σ_ϵ are taken to be Jeffrey's priors, so that $\pi(\Phi, \Sigma_\epsilon) \propto 1/|\Sigma_\epsilon|^2$. For Model BV2, we assume a $N(\mu_{\phi_0}, \sigma_{\phi_0}^2)$ prior for ϕ_0, a $N(\mu_{\phi_1}, \sigma_{\phi_1}^2)$ prior for ϕ_1, and an inverse Gamma IG(s_1, s_2) prior for σ^2. The joint posterior distribution of the parameters, the complete conditional distributions and the sampling algorithms under the constant proportion and the stochastic proportion models are derived in Sections 3.1 and 3.2 below.

3.1. Posterior Analysis under Constant Proportion Models

Under Model MV1, the complete conditional distributions for the parameters describing the Weibull NHPP growth curve are proportional to the joint posterior, and are

$$\begin{aligned}
\log f(c|\cdots) &\propto -\sum_{j=1}^{T} m_{t_j} + \sum_{j=1}^{T} Y_{t_j} \log(m_{t_j}) - \frac{1}{2\sigma_c^2}(c - \mu_c)^2, \\
\log f(d|\cdots) &\propto -\sum_{j=1}^{T} m_{t_j} + \sum_{j=1}^{T} Y_{t_j} \log(m_{t_j}) - \frac{1}{2\sigma_d^2}(d - \mu_d)^2, \text{ and} \\
\log f(e|\cdots) &\propto -\sum_{j=1}^{T} m_{t_j} + \sum_{j=1}^{T} Y_{t_j} \log(m_{t_j}) - \frac{1}{2\sigma_e^2}(e - \mu_e)^2 \qquad (15)
\end{aligned}$$

The complete conditional distribution of $\mathbf{p}$ is a Dirichlet $D_K(Y^{(1)}+\frac{1}{2},\cdots,Y^{(K)}+\frac{1}{2})$ distribution. Under Model BV1, the complete conditional distributions for c, d, and e are given by (15), while that of p is the Beta$(Y^{(1)}+\frac{1}{2}, Y^{(2)}+\frac{1}{2})$ distribution.

3.2. Posterior Analysis under Stochastic Proportion Models

For Model MV2, sampling of Φ and Σ_ϵ within each Gibbs iteration proceeds as follows (see Sun and Ni [13]). Let $\hat{\Phi}_M = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Q}$, $\hat{\Sigma}_\epsilon = S(\hat{\Phi}_M)/T$, where $S(\hat{\Phi}_M) = (\mathbf{Q}-\mathbf{X}\hat{\Phi}_M)'(\mathbf{Q}-\mathbf{X}\hat{\Phi}_M)$. The posterior of Φ is then $\mathrm{N}(\hat{\Phi}_M, \hat{\Sigma}_\epsilon \bigotimes (\mathbf{X}'\mathbf{X})^{-1})$. Letting $\Sigma_\epsilon = \{\sigma_{ij}\}$, and $S(\hat{\Phi}_M) = \{s_{ij}\}$, the posterior distribution of σ_{ii} is inverse Gamma IG$(T-2K, s_{ii})$.

Under Model BV2, the complete conditional distributions for c, d, and e are given by (15). The complete conditional distribution of the latent vector $\mathbf{q}$ is

$$\begin{aligned}\log f(\mathbf{q}|\cdots) &\propto \sum_{j=1}^{T} y_{t_j}^{(1)} q_{t_j} - \sum_{j=1}^{T} Y_{t_j}\log(1+\exp(q_{t_j})) - \frac{1}{2\sigma^2}(1-\phi_1^2)(q_{t_1}-\frac{\phi_0}{1-\phi_1})^2 \\ &- \frac{1}{2\sigma^2}\sum_{j=1}^{T-1}(q_{t_{j+1}}-\phi_0-\phi_1 q_{t_j})^2 \end{aligned} \tag{16}$$

which is dominated by a multivariate normal density. The Metropolis-Hastings algorithm is used to generate samples of $\mathbf{q}$ from its complete conditional distribution using a multivariate normal proposal. The complete conditional distribution of σ^2 is an inverse gamma $IG(a_1, a_2)$ where

$$\begin{aligned} a_1 &= s_1 + \frac{1}{2}T - 1, \text{ and} \\ a_2 &= s_2 + \frac{1}{2}\{(1-\phi_1^2)(q_{t_1}-\frac{\phi_0}{1-\phi_1})^2 + \sum_{j=1}^{T-1}(q_{t_{j+1}}-\phi_0-\phi_1 q_{t_j})^2\}. \end{aligned} \tag{17}$$

The complete conditional distribution of ϕ_0 is a normal distribution with mean and variance

$$\begin{aligned} \text{Mean} &= \frac{\sigma_{\phi_0}^2\{q_{t_1}(1+\phi_1)+\sum_{j=1}^{T-1}(q_{t_{j+1}}-\phi_1 q_{t_j})\}+\sigma^2\mu_{\phi_0}}{\sigma^2+\sigma_{\phi_0}^2(\frac{1+\phi_1}{1-\phi_1}+T-1)} \\ \text{Variance} &= \frac{\sigma^2\sigma_{\phi_0}^2}{\sigma^2+\sigma_{\phi_0}^2(\frac{1+\phi_1}{1-\phi_1}+T-1)} \end{aligned} \tag{18}$$

where μ_{ϕ_0} and $\sigma_{\phi_0}^2$ are respectively the prior mean and variance for ϕ_0. The complete conditional distribution for ϕ_1 is the normal distribution with mean

$$\text{Mean } = \widetilde{\sigma}^2\left(\frac{\sum_{j=1}^{T-1}(q_{t_{j+1}}-\phi_0)q_{t_j}}{\sigma^2}+\frac{\mu_{\phi_1}}{\sigma_{\phi_1}^2}\right)$$

where the variance $\widetilde{\sigma}^2$ is given by

$$\text{Variance } = \widetilde{\sigma}^2 = \frac{\sigma^2 \sigma_{\phi_1}^2}{\sigma^2 + \sigma_{\phi_1}^2} \sum_{j=1}^{T-1} q_{t_j}^2$$

where μ_{ϕ_1} and $\sigma_{\phi_1}^2$ are the prior mean and variance for ϕ_1. Straightforward draws for ϕ_1 can be made subject to the restriction that $|\phi_1| < 1$. Chib and Albert [16] provides a theoretical justification for this sampling strategy.

Alternatively, using noninformative priors for ϕ_0 and ϕ_1, the complete conditional distributions have the following forms. The $f(\phi_0|\cdots)$ is a normal distribution with mean and variance given respectively by

$$\frac{q_{t_1}(1+\phi_1) + \sum_{j=1}^{T-1}(q_{t_{j+1}} - \phi_1 q_{t_j})}{\frac{1+\phi_1}{1-\phi_1} + T - 1}$$

$$\frac{\sigma^2}{\frac{1+\phi_1}{1-\phi_1} + T - 1}$$

Subject to the restriction $|\phi_1| < 1$, draws for ϕ_1 are made from a normal distribution with mean and variance

$$\frac{\sum_{j=1}^{T-1} q_{t_j}(q_{t_{j+1}} - \phi_0)}{\sum_{j=1}^{T-1} q_{t_j}^2}$$

$$\frac{\sigma^2}{\sum_{j=1}^{T-1} q_{t_j}^2}$$

3.3. Prediction and Model Selection

Prediction and model selection are based on the predictive density. In general, if z denotes a new observation and $\mathbf{Y}$ denotes the given data, the posterior predictive density has the form

$$f(z|\mathbf{Y}) = \int f(z|\Psi) f(\Psi|\mathbf{Y}) d\Psi \tag{19}$$

In cases when it is cumbersome to obtain an analytical form for this predictive density, it is straightforward to compute it based on the converged MCMC samples $\{\Psi_j^*, j = 1, \cdots, k\}$ from the posterior distribution of the parameters (see Gelfand [17]). Samples $\{Y_j^*, j = 1, \cdots, k\}$ from the posterior predictive density are obtained from $f(Y|\Psi_j^*)$. Marginally Y_j^* is a sample from $f(Y|Y_{obs})$. The ith element $y_{i,j}^*$ of Y_j^* is then a sample from $f(y_i|Y_{obs})$.

In our analysis, we have used $k = 2500$ iterations to compute predictive distributions. For each iteration, we first generate a Poisson observation and then generate categorized defects under

each model, giving the predictive samples. The 50% and 95% credible sets for the prediction are based on the (25th, 75th) and (2.5th, and 97.5th) percentiles of this sample, respectively.

We address the question of model comparison via the L-measure (Ibrahim, Chen and Sinha [18]). In general, given the posterior distribution π^* based on data $\mathbf{Y} = (Y_1, \cdots, Y_n)^*$ and unknown parameters Ψ, and samples Z_i from the posterior predictive density, the L-measure is defined as

$$L(\pi^*) = \sum_{i=1}^{n} \mathrm{Var}(Z_i|\mathbf{Y}) + \frac{w}{1+w} \sum_{i=1}^{n} (\mu_i - Y_i)^2 \tag{20}$$

where $\mu_i = \mathrm{E}^{\pi^*}(\mathrm{E}[Z_i|\Psi])$, and w is a weighting constant usually taken as 1. The L-measure is a sum of the predictive variance and a weighted bias term. Small values imply a good model, and computation given MCMC output is straightforward.

4. Simulation Study

In this section, we investigate the efficacy of the proposed methods using simulated data. Under each of models BV1, MV1, BV2, and MV2, we simulate 50 weeks of data from a multinomial NHPP process using the parameter values listed in the last column of Tables 4.1 - 4.4, respectively. We use the first 40 weeks of data for model fitting, leaving the last 10 for predictive cross-validation.

Here, and in Section 5, we choose parameters of the prior distributions in a "semi-empirical" fashion. Since it is likely that there will be few remaining undetected defects at a later stage of testing, we take 120% of the sum of the defects for the fitting portion to serve as the mean of the prior for θ. The prior for β is taken to be a small value, either 0.05 or 0.01, to indicate the belief that the defect discovery rate is relatively small. The prior for α is set to be 1, essentially assuming that the exponential model will be sufficient. For the AR(1) parameters under Model BV2, we fit an AR(1) model to the empirical proportions and use the resulting estimates to provide prior specifications for ϕ_0, ϕ_1, and σ^2. The prior variances are chosen to be sufficiently large to correspond to noninformative prior specifications. Specifically, the priors under Model BV1 for c, d and e are respectively $N(5.027, 100)$, $N(-4, 1)$, and $N(1, 1)$. Under Model MV1, the priors for c, d and e are respectively $N(5.026, 100)$, $N(-4, 1)$ and $N(1, 1)$. Under Model BV2, the prior on c is $N(4.944, 100)$, d is $N(-4, 1)$, e is $N(1, 1)$, ϕ_0 is $N(-1, 1)$, ϕ_1 is $N(0.9, 0.04)$, and σ^2 is Inverse Gamma $IG(10, 2)$. Under Model MV2, the priors for c, d, and e are respectively $N(5.848, 100)$, $N(-4, 1)$ and $N(1, 1)$, while Jeffrey's priors are employed for Φ and Σ.

We run 5000 iterations of the Gibbs sampler, with inference based on the last 2500 iterations. Tables 4.1-4.4 present the posterior summaries, including the mean, standard deviation and the 95% credible interval for the parameters under each model.

—- Table 4.1 about here —-

—- Table 4.2 about here —-

—- Table 4.3 about here —-

—- Table 4.4 about here —-

In all cases, the fit gives credible intervals containing the true parameter values, indicating the feasibility of the proposed methods even for small data sets. Prior sensitivity studies revealed that different choices of the prior means and the prior variances yield stable posterior estimates as long as the prior variances are sufficiently large (diffuse prior information). Convergence was monitored via BOA function suites (see [19]).

In the next section, we present an application to defects discovered during the System Test phase of a large operating system development project and categorized according to type.

5. Illustration for Categorized Software Defects from an IBM Software Development Project

5.1 Data Description

For illustration, we fit our model to defects discovered over a 45 week System Test period for a large operating system software component, where the defects were classified using the Orthogonal Defect Classification (ODC) scheme introduced in Chillarege *et al.* [6]. ODC identifies seven distinct defect types for standard Requirements/Design/Code activities, which include System Test. These are Assignment/Initialization, Checking, Algorithm/Method, Function/Class/Object, Timing/Serialization, Interface/O-O Messages, and Relationship. Here, we focus on defect types Assignment/Initiation (AI), Checking (CH), Algorithm/Method (AG), and Function/Class/Object(FC) only. Assignment/Initialization defects can briefly be described as ones in which an incorrect of missing assignment is the cause of the error, while checking defects are due to an incorrect or missing check in the code. Algorithm/Method defects are attributable to an incorrect or missing algorithm step, while Function/Class/Object defects indicate incorrect or missing functionality relative to requirements. See Chillarege *et al.* [6] for more precise definitions of each type. The paper of Ray *et al.* [8] investigated the relationship between AI and CH defects, finding that discovery of CH defects is often predicated on discovery and removal of initialization defects. Here, we investigate dependencies between all four defect types, with investigation of AI and CH defects only used to illustrate the bivariate model.

In our example, we have a total of 122 AI defects, 59 CH defects, 91 AG defects, and 28 FC defects, for a total of 300. Figure 5.1 shows the cumulative number of defects each week for the four types of defects. We use data from the first 35 weeks for fitting, leaving the last 10 weeks for predictive evaluation. The prior parameters are obtained as described in Section 4. Posterior predictions are obtained following the method discussed in Section 3.3. We choose to explicitly model AI, CH, and AG defects, letting the proportion of FC defects (and the dependence of the AI, CH, and AG proportions on FC defects) be defined implicitly through the multinomial proportion relationship.

—- Figure 5.1 about here —-

5.2 Estimation Results

Results from fitting Models MV1, BV1, MV2 and BV2 to the data are shown in Tables 5.1 - 5.4, respectively. For Models BV1 and BV2, we fit to AI and CH defects only, with the total defects the sum of AI and CH defects only. For Models MV1 and MV2, we fit to all four defect types.

Several observations can be drawn from the results. First, for each of the fitted models, the evolution of the total number of defects follows a Weibull model instead of exponential, as $\log(\alpha)$ is significantly larger than zero. The large value of α indicates a rapidly increasing hazard rate, *i.e.* the rate at which defects are found increases quickly as ST progresses. Since the models are additive for the total number of defects and the underlying type probabilities, it is not surprising to find that the posterior estimates for the Weibull parameters, θ, β, and α are almost the same for MV1 and MV2 and for BV1 and BV2. The top and bottom left-hand plots in Figure 5.2 show that the posterior predictions for the evolution of the cumulative total defects are slightly larger than the actual, perhaps due to the use of a prior mean for c that is 20% larger than the observed total up through 35 weeks.

Turning now to models for the individual defect type proportions, Table 5.1 shows results for the constant proportion model MV1, while Table 5.3 shows results for the stochastic proportion model MV2. From Table 5.3, we see little evidence of dependence between the log odds of AI defects and the previous log odds for AI and CH defects (small values of ϕ_{11} and ϕ_{22}), and negative correlation with the log odds of AG defects. In other words, the higher the log odds of AG defects found in the previous period, the lower the log odds of finding AI defects in the current time period. On the other hand, the significantly positive values for ϕ_{21}, ϕ_{22}, and ϕ_{23} indicate that the log odds of CH defects is strongly dependent on the log odds of AI defects in the previous time period ($\phi_{21} = 1.064$), as well as moderately dependent on the log odds of CH and AG defects in the previous period ($\phi_{22} = 0.304$ and $\phi_{23} = 0.422$). This agrees with previous findings of Ray *et al.* [8], that many CH defects cannot be discovered until AI defects are detected and removed. The log odds for AG defects tends to show little dependence on the log odds for previously detected AI and CH defects, and strong negative dependence on the log odds of AG defects in the previous time period, as evidenced by estimated autoregressive parameters $\phi_{31} = 0.126, \phi_{32} = 0.128$ and $\phi_{33} = -0.688$ respectively.

When considering only the relationship between AI and CH defects, captured in the BV2 model, the stochastic model results shown in Table 5.4 indicate little dependence between the log odds of AI defects and the log odds of AI defects in the previous time period. However, the estimated parameters are much more variable than those obtained using MV2. The large estimated value of σ^2 suggests considerable uncertainty in the evolution of the log odds over time, especially in comparison to the σ_1^2 value for MV2. This may be due to the fact that the BV model fails to allow for possible relations with other defect types.

—- Table 5.1 about here —

—- Table 5.2 about here —

—- Table 5.3 about here —

—- Table 5.4 about here —

The top and bottom right-hand plots in Figure 5.2 and the left- and right-hand plots in Figure 5.3 show the predicted number of defects by type for the constant and stochastic proportion models, respectively. For Model MV1, the posterior predicted cumulative mean for each type of defect is proportional to the percentage of observed defects. Given the fact that the predicted overall mean is higher than the observed total, we see that all types are somewhat over-predicted. Furthermore, the over-prediction is proportional to the percentages. For instance, AI defects make up about 40% of the total defects. As a result, over-prediction is worst for AI defects.

The MV2 model provides more accurate predictions for AI and AG defects than the MV1 model, but much worse for CH and FC defects. The number of CH defects is grossly overestimated using MV2. The predictions shown in Figure 5.3 for AI and CH defects are quite similar using BV1 and BV2. Comparison of the constant and stochastic proportion models based on L-measures (Table 5.5) indicates that the constant proportion models are to be preferred for this data set, based on the overall predictive ability of the models.

Figure 5.3 also compares the posterior predictions for AI and CH defects with those obtained from the dynamic GC model using covariates, as reported in Ray *et al.* [9]. The dynamic growth curve (DGC) approach models each category of defects separately, using previously observed defect volumes in other categories as covariates. In the MV proportion approach, the overall mean, discovery rate β, and usage rate, α, are modeled for all types combined, independent of the distribution of defects among categories. In the DGC approach, these parameters are modeled for each category separately, and can all potentially be dependent on significant covariates. From Figure 5.3, Models BV1 and BV2 appear to provide improved predictions over those obtained using the DGC approach.

—- Figure 5.2 about here —

—- Figure 5.3 about here —

—-Table 5.5 about here —-

6. Conclusion

In this paper we have proposed a general class of models for characterizing the evolution of K categories of event data over time, both for $K = 2$ and more generally for $K > 2$. Bayesian inference procedures were outlined in detail and simulation investigations indicated the feasibility of the methods. The model framework was found to be useful for characterizing software reliability based on the defect discovery process for defects of different types. The results of the particular application reported here suggest that there are significant interactions between different types of defects. Information on these dependencies can provide additional insight into the progress of the development process and aid decision making regarding readiness for release.

The proposed models could be further developed by incorporating prior release information into the modeling, *e.g.* in the formulation of priors. Another potential avenue for development is the incorporation of exogenous predictors into the autoregressive framework. We leave these investigations for future research.

References

[1] Jelinski Z, Moranda P. Software Reliability Research. In Statistical Computer Performance Evaluation, Freiberger W (ed). Academic Press: New York, 1972.

[2] Goel AL, Okumoto K. Time-dependent error-detection rate model for software reliability and other performance measures. *IEEE Trans. Rel.* 1979; R-28(1): 206-211.

[3] Yamada S, Ohba M, Osaki S. S-shaped reliability growth modeling for software error detection. *IEEE Transactions on Reliability*, 1983; 32, 475-478.

[4] Kuo L, Yang T. Bayesian computation for nonhomogeneous Poisson processes in software reliability. *J. Amer. Statist. Assoc.* 1996; 91: 763-773.

[5] Singpurwalla N, Wilson S. Statistical Methods in Software Engineering: Reliablity and Risk. 1999; Springer: New York.

[6] Chillarege R, Bhandari I, Chaar J, Halliday M, Moebus D, Ray B, and Wong M. Orthogonal defect classification-a concept for in-process measurement. *IEEE Trans. Software Eng.* 1992; 18: 943-956.

[7] Chillarege R, and Biryani S. Identifying risk using ODC based growth curve models. *IEEE Trans. Software Eng.* 1994; 282-288.

[8] Ray B, Bhandari I, Chillarege R. Reliability growth for typed defects. In *Proc. IEEE Annual Reliability and Maintainability Symposium* 1992; 327-336.

[9] Ray B, Liu Z, Ravishanker, N. Dynamic reliability models for software using time-dependent covariates. Technical Report, Department of Statistics, University of Connecticut, 2004.

[10] Singpurwalla N, Soyer R. Assessing (software) reliablity growth using a random coefficient autoregressive process and its ramifications. *IEEE Trans. on Software Eng.* 1985; SE-11 12: 1456-1464.

[11] Kenney G. Estimating defects in commercial software during operational use. *IEEE Trans. Rel.* 1993; 42: 107-115.

[12] Ross S. An Introduction to Probability Models. 2004; Academic Press: New York.

[13] Sun D, Ni S. Bayesian analysis of vector-autoregressive models with noninformative priors. *J. Statist. Planning and Inference* 2004; 121: 291-309.

[14] Gelfand AE, Smith AFM. Sampling based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* 1990; 85: 398-409.

[15] Robert C, Casella G. Monte Carlo Statistical Methods. 1999; Springer: New York.

[16] Chib S, Albert JH. Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. *J. Bus. Econ. Statist.* 1993; 11: 1-15.

[17] Gelfand AE. Model Determination using sampling-based methods. In Markov chain Monte Carlo in Practice, Gilks WR, Richardson S, Spiegelhalter DJ (eds). Chapman and Hall: London, 1995; 145-161.

[18] Ibrahim JG, Chen MH, Sinha D. Criterion based methods for Bayesian model assessment. *Statistica Sinica.* 2001; 11: 419-443.

[19] Bayesian Output Analysis. http://www.public-health.uiowa.edu/boa/

Table 4.1. Posterior Summary for Data Simulated under Model BV1

Parameter	Mean	Std. Dev	95% Credible Interval	True Parameter
$\log(\theta)$	5.066	0.345	(4.751,6.425)	5
$\log(\beta)$	-3.188	0.326	(-4.293,-2.717)	-3
$\log(\alpha)$	0.027	0.116	(-0.288, 0.207)	0
p	0.583	0.042	(0.499, 0.663)	0.6

Table 4.2. Posterior Summary for Data Simulated under Model MV1

Parameter	Mean	Std. Dev	95% Credible Interval	True Parameter
$\log(\theta)$	5.175	0.322	(4.816,6.167)	5
$\log(\beta)$	-3.574	0.321	(-4.337,-2.995)	-3
$\log(\alpha)$	0.046	0.136	(-0.257,0.287)	0
p_1	0.431	0.043	(0.349,0.511)	0.4
p_2	0.283	0.040	(0.208,0.364)	0.3
p_3	0.205	0.036	(0.138,0.280)	0.2
p_4	0.081	0.024	(0.039,0.136)	0.1

Table 4.3. Posterior Summary for Data Simulated under Model BV2

Parameter	Mean	Std. Dev	95% Credible Interval	True Parameter
$\log(\theta)$	4.904	0.170	(4.676,5.324)	5
$\log(\beta)$	-2.819	0.248	(-3.319,-2.353)	-3
$\log(\alpha)$	-0.035	0.113	(-0.297, 0.157)	0
ϕ_0	-0.178	0.118	(-0.434, 0.024)	-0.2
ϕ_1	0.643	0.150	(0.330, 0.908)	0.6
σ^2	0.247	0.094	(0.129, 0.484)	0.25

Table 4.4. Posterior Summary for Data Simulated under Model MV2

Parameter	Mean	Std. Dev	95% Credible Interval	True Parameter
$\log(\theta)$	5.877	0.106	(5.703, 6.135)	5.75
$\log(\beta)$	-7.137	0.384	(-7.857,-6.355)	-7.5
$\log(\alpha)$	0.728	0.069	(0.569, 0.849)	0.8
Φ_{01}	0.837	0.351	(0.103, 1.602)	1.179
Φ_{02}	0.060	0.148	(-0.217, 0.394)	0.693
Φ_{03}	0.179	0.318	(-0.530, 0.672)	0.916
σ_1^2	0.765	0.358	(0.229, 1.547)	0.250
σ_2^2	0.148	0.097	(0.025, 0.384)	0.250
σ_3^2	0.711	0.563	(0.162, 2.283)	0.250
Φ_{111}	0.086	0.169	(-0.231, 0.471)	0.400
Φ_{112}	0.094	0.162	(-0.178, 0.479)	0
Φ_{113}	0.108	0.198	(-0.194, 0.591)	0
Φ_{121}	0.186	0.382	(-0.538, 0.811)	0
Φ_{122}	0.147	0.244	(-0.351, 0.536)	0.400
Φ_{123}	0.092	0.344	(-0.629, 0.732)	0
Φ_{131}	0.043	0.202	(-0.367, 0.444)	0
Φ_{132}	0.136	0.242	(-0.385, 0.524)	0
Φ_{133}	0.089	0.267	(-0.478, 0.569)	0.500

Table 5.1. Posterior Summary for IBM Categorized Defects under Model MV1

Parameter	Mean	Std. Dev	95%Credible Interval
$\log(\theta)$	5.805	0.027	(5.756,5.856)
$\log(\beta)$	-9.814	0.500	(-10.766,-8.816)
$\log(\alpha)$	1.146	0.051	(1.0414,1.238)
$p^{(1)}$	0.4135	0.029	(0.357,0.470)
$p^{(2)}$	0.1888	0.023	(0.144,0.234)
$p^{(3)}$	0.3025	0.027	(0.252,0.354)
$p^{(4)}$	0.0952	0.017	(0.064,0.132)

Table 5.2. Posterior Summary for IBM Categorized Defects under Model BV1

Parameter	Mean	Std. Dev	95% Credible Interval
$\log(\theta)$	5.302	0.051	(5.214,5.411)
$\log(\beta)$	-9.08	0.528	(-10.144,-8.103)
$\log(\alpha)$	1.053	0.060	(0.940, 1.171)
$p^{(1)}$	0.686	0.035	(0.618, 0.753)
$p^{(2)}$	0.314	0.035	(0.246, 0.385)

Table 5.3. Posterior Summary for IBM Categorized Defects under Model MV2

Parameter	Mean	Std. Dev	95%Credible Interval
$\log(\theta)$	5.805	0.026	(5.755, 5.857)
$\log(\beta)$	-9.810	0.496	(-10.790,-8.859)
$\log(\alpha)$	1.146	0.050	(1.046, 1.242)
Φ_{01}	3.007	0.558	(2.120,4.330)
Φ_{02}	1.431	0.264	(1.010, 2.044)
Φ_{03}	3.848	0.743	(2.657, 5.534)
σ_1^2	0.648	0.186	(0.355, 1.156)
σ_2^2	0.314	0.058	(0.133, 0.389)
σ_3^2	1.533	0.125	(1.367, 1.898)
Φ_{111}	0.129	0.030	(0.072, 0.197)
Φ_{112}	-0.072	0.064	(-0.195, 0.057)
Φ_{113}	-0.325	0.028	(-0.361,-0.242)
Φ_{121}	1.064	0.057	(0.895, 1.144)
Φ_{122}	0.304	0.020	(0.275, 0.365)
Φ_{123}	0.422	0.167	(0.087, 0.744)
Φ_{131}	0.126	0.062	(-0.048, 0.217)
Φ_{132}	0.128	0.094	(-0.068, 0.310)
Φ_{133}	-0.688	0.035	(-0.755, -0.617)

Table 5.4. Posterior Summary for IBM Categorized Defects under Model BV2

Parameter	Mean	Std. Dev	95%Credible Interval
$\log(\theta)$	5.288	0.040	(5.2075,5.366)
$\log(\beta)$	-8.970	0.623	(-10.189, -7.735)
$\log(\alpha)$	1.040	0.072	(0.887, 1.173)
ϕ_0	1.003	0.493	(0.097, 2.026)
ϕ_1	0.079	0.361	(-0.664,0.6946)
σ^2	2.604	1.088	(1.158, 5.360)

Table 5.5. *L*-Measure for Model Comparison

	Models	L-Measure
1	Model BV1	439.6
2	Model BV2	560.7
3	Model MV1	1288.7
4	Model MV2	1663.9

Figure 5.1. Cumulative Categorized Defects

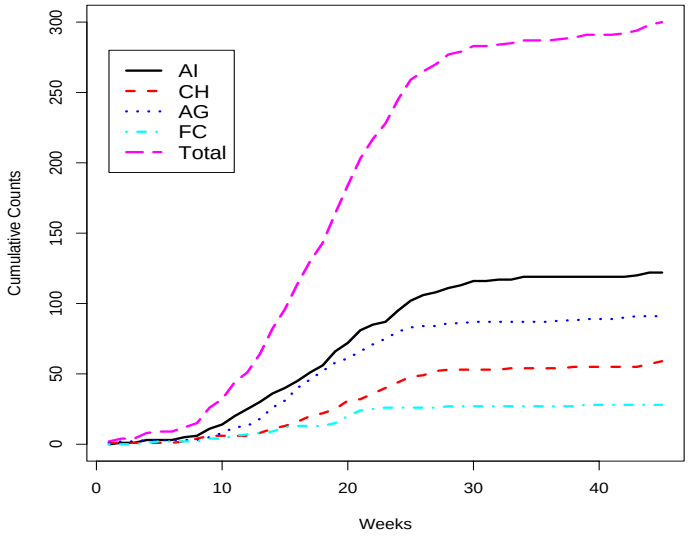

Figure 5.2. Posterior Predictions for Models MV1 and MV2. Top left: Cumulative total predicted defects using MV1; Top right: Cumulative total predicted defects by type using MV1; Bottom left: Cumulative total predicted defects using MV2; Bottom left: Cumulative total predicted defects by type using MV2.

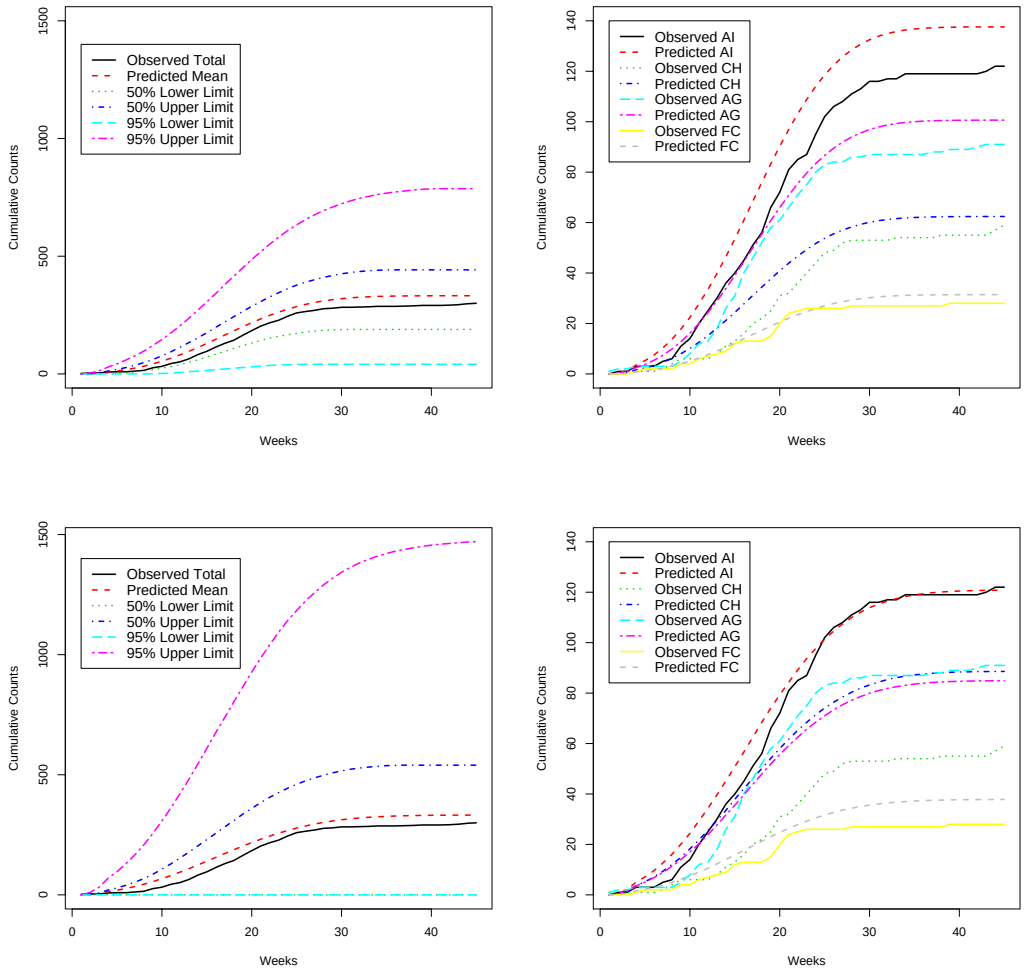

Figure 5.3. Posterior Predictions for Models BV1 (left) and BV2 (right), along with predictions from Dynamic Growth Curve models of Ray *et al.* (2004).