Bayesian Analysis For Partially Complete Time and Type of Failure Data

Debasis Kundu

Department of Mathematics & Statistics
Indian Institute of Technology Kanpur

April 24, 2013

Outline

1. Competing Risks
2. Different Approaches
3. Model Formulation
4. Model and Prior Assumptions
5. Posterior Analysis
6. Bayes Estimates
7. Data Analysis
8. Analysis in Presence of Covariates

Outline

1. Competing Risks
2. Different Approaches
3. Model Formulation
4. Model and Prior Assumptions
5. Posterior Analysis
6. Bayes Estimates
7. Data Analysis
8. Analysis in Presence of Covariates

What is Competing Risk?

In medical studies or in reliability analysis, an investigator is often interested in the assessment of a specific risk factor in the presence of other risks. It is well known as the competing risks problem, in the statistical literature. Usually the competing risks data consists of failure time and an indicator denoting the cause of failure.

Example:

1. A computer monitor may fail due to different reason, temperature, high voltage, improper handling etc.
2. A car engine may fail due to different reasons.
3. A patient can die due to different reasons, for example heart disease, kidney failure etc.

The main aim is to analyze the lifetime distribution of a particular cause in presence of other risk factors.

Associated Issues

1. Typically competing risk data looks like (T, Δ). Here T denotes the lifetime of the item and Δ denotes the cause of failure.
2. The life time distributions associated with different causes of failure may be assumed to be dependent or independent.
3. Although the assumption of dependence seems more reasonable, there is some identifiability issue.
4. It is not possible to test the assumption of independence of the failure time distributions without the presence of covariates.

Data in our Experiment

In our experiment we may observe the following four types of observations:

1. Time and cause of failure both are known: $(T = t, \Delta = j)$
2. Time is known, cause of failure is unknown: $(T = t, \Delta = *)$.
3. Time censored, cause of failure is unknown: $(T > t, \Delta = *)$.
4. Time censored, cause of failure is known: $(T > t, \Delta = j)$.

Outline

1. Competing Risks
2. Different Approaches
3. Model Formulation
4. Model and Prior Assumptions
5. Posterior Analysis
6. Bayes Estimates
7. Data Analysis
8. Analysis in Presence of Covariates

Latent failure time modeling

Cox (1959, JRSS B) proposed the latent failure time modeling to analyze competing risk data. The latent failure time model has the following form:

$$T = \min\{T_1, \cdots, T_M\}$$

here T denotes the observed failure time of an item, and $T_1, \cdots, T_M$ are the latent failure times of the M different causes, and they are assumed to be independent.

Cause specific hazard

Prentice et al. (1978, Biometrics) proposed the cause specific hazard model to analyze competing risks data.
In this case the overall hazard rate is defines as usual:

$$\lambda(t) = \lim_{dt \to 0} \frac{P(t \leq T < t + dt | T \geq t)}{dt}$$

The cause specific hazard rate representing the instantaneous risk of dying of cause j:

$$\lambda_j(t) = \lim_{dt \to 0} \frac{P(t \leq T < t + dt, \Delta = j | T \geq t)}{dt}.$$

By the law of total probability

$$\lambda(t) = \sum_{j=1}^{M} \lambda_j(t).$$

Outline

1. Competing Risks
2. Different Approaches
3. Model Formulation
4. Model and Prior Assumptions
5. Posterior Analysis
6. Bayes Estimates
7. Data Analysis
8. Analysis in Presence of Covariates

Data and Likelihood Contribution

It is assumed that we have only two possible causes of failures. Hence we have the following types of observations:

$$(t, 1), \quad (t, 2), \quad (t, *), \quad (t*, *), \quad (t*, 1), \quad (t*, 2).$$

Therefore, likelihood contribution from the observations from different sets are as follows:

$$f_1(t)S_2(t), \quad f_2(t)S_1(t), \quad f(t), \quad S(t)$$

$$\int_t^\infty f_1(y)S_2(y)dy, \quad \int_t^\infty f_2(y)S_1(y)dy.$$

Data and Likelihood Contribution: Covariates

Each item might have some associated covariates also. Hence we have the following types of observations:

$$(t, 1, x), \quad (t, 2, x), \quad (t, *, x), \quad (t*, *, x), \quad (t*, 1, x), \quad (t*, 2, x).$$

Therefore, likelihood contribution from the observations from different sets are as follows:

$$f_1(t; x)S_2(t; x), \quad f_2(t; x)S_1(t; x), \quad f(t; x), \quad S(t; x)$$

$$\int_t^\infty f_1(y : x)S_2(y; x)dy, \quad \int_t^\infty f_2(y; x)S_1(y; x)dy.$$

Existing Work

1. Non-parametric work has been first done by Dinse.
2. Parametric inference under the frequentist set up has been performed using exponential and Weibull latent failure distribution assumptions.

All the results are asymptotic in nature. Our main aim is to consider Bayesian inference under quite a flexible priors.

Outline

1. Competing Risks
2. Different Approaches
3. Model Formulation
4. Model and Prior Assumptions
5. Posterior Analysis
6. Bayes Estimates
7. Data Analysis
8. Analysis in Presence of Covariates

Lifetime distribution

Here it is assumed that $M = 2$. The latent failure times have distributions with the same shape parameter but different scale parameters, *i.e.* $X_1 \sim \mathrm{WE}(\alpha, \lambda_1)$ and $X_2 \sim \mathrm{WE}(\alpha, \lambda_2)$ and they are independent. $\mathrm{WE}(\alpha, \lambda)$ has the PDF

$$f(x; \alpha, \lambda) = \alpha \lambda x^{\alpha - 1} e^{-\lambda x^{\alpha}}; \quad x > 0.$$

It is well known that Weibull distribution is a very flexible lifetime distribution. Moreover, $\min\{X_1, X_2\} \sim \mathrm{WE}(\alpha, \lambda)$, $\lambda = \lambda_1 + \lambda_2$.

Prior Assumptions on λ_1 and λ_2

$$\pi_0(\lambda|a_0, b_0) = \frac{b_0^{a_0}}{\Gamma(a_0)}\lambda^{a_0-1}e^{-b_0\lambda}$$

$$\pi(\lambda_1/\lambda|a_1, a_2) = \frac{\Gamma(a_1+a_2)}{\Gamma(a_1)\Gamma(a_2)}\left(\frac{\lambda_1}{\lambda}\right)^{a_1-1}\left(1-\frac{\lambda_1}{\lambda}\right)^{a_2-1}$$

$$\begin{aligned}\pi(\lambda_1, \lambda_2|a_0, b_0, a_1, a_2) &= \frac{\Gamma(a_1+a_2)}{\Gamma(a_0)}(b_0\lambda)^{a_0-a_1-a_2} \times \\ &\frac{b_0^{a_1}}{\Gamma(a_1)}\lambda_1^{a_1-1}e^{-b_0\lambda_1} \\ &\times\frac{b_0^{a_2}}{\Gamma(a_2)}\lambda_2^{a_2-1}e^{-b_0\lambda_2}\end{aligned}$$

Beta-Gamma Prior

1. $$\pi(\lambda_1, \lambda_2 | a_0, b_0, a_1, a_2) \sim \mathrm{BG}(b_0, a_0, a_1, a_2)$$

2. If $a_0 = a_1 + a_2$, λ_1 and λ_2 are independent

3. λ_1 and λ_2 positively or negatively correlated if $a_0 > a_1 + a_2$ or $a_0 < a_1 + a_2$ respectively. When $a_0 = a_1 + a_2$, λ_1 and λ_2 become independent.

Mean and Variance of the Beta-Gamma Prior

If $(\lambda_1, \lambda_2) \sim \mathrm{BG}(b_0, a_0, a_1, a_2)$, then for $i = 1, 2$

$$
\begin{aligned}
E(\lambda_i) &= \frac{a_0 a_i}{b_0(a_1 + a_2)} \\
V(\lambda_i) &= \frac{a_0 a_i}{b_0^2(a_1 + a_2)} \times \left\{ \frac{(a_i + 1)(a_0 + 1)}{a_1 + a_2 + 1} - \frac{a_0 a_i}{a_1 + a_2} \right\}
\end{aligned}
$$

Prior Assumption on α

No specific prior has been assumed on α. It is assumed that the prior on α has a support on $(0, \infty)$, the PDF is log-concave. It is independent of the prior on (λ_1, λ_2).
Note that several standard distribution functions have log-concave PDF, for example (i) log-normal, (ii) gamma (shape parameter greater than 1), (iii) Weibull (shape parameter greater than 1), (iii) generalized exponential distribution (shape parameter greater than 1) etc.

Outline

1. Competing Risks
2. Different Approaches
3. Model Formulation
4. Model and Prior Assumptions
5. Posterior Analysis
6. Bayes Estimates
7. Data Analysis
8. Analysis in Presence of Covariates

Shape Parameter Known

In this case

$$\pi(\lambda_1, \lambda_2 | Data) \sim \mathrm{BG}\left(B_0, A_0, A_1, A_2\right),$$

here

$$\begin{aligned} B_0 &= b_0 + T(\alpha) \\ A_0 &= a_0 + r_1 + r_2 + r_3 \\ A_1 &= a_1 + r_1 + r_4 \\ A_2 &= a_2 + r_2 + r_5. \end{aligned}$$

$$T(\alpha) = \sum_{i=1}^{n} t_i^{\alpha}$$

Outline

1. Competing Risks
2. Different Approaches
3. Model Formulation
4. Model and Prior Assumptions
5. Posterior Analysis
6. Bayes Estimates
7. Data Analysis
8. Analysis in Presence of Covariates

Bayes Estimates: Squared Error Loss Function

The Bayes estimates of λ_1 and λ_2 under Squared error loss function becomes

$$\widehat{\lambda}_{1B}(\alpha) = \frac{(n + a_0)(r_1 + r_4 + a_1)}{(b_0 + T(\alpha))(n_1 + r_1 + a_1 + a_2)}$$

$$\widehat{\lambda}_{2B}(\alpha) = \frac{(n + a_0)(r_2 + r_5 + a_2)}{(b_0 + T(\alpha))(n_1 + r_1 + a_1 + a_2)}$$

We have explicit expressions for posterior variance also. The credible interval for λ_1 and λ_2 can be obtained numerically.

Bayes Estimate and MLE

The Bayes estimates of λ_1 and λ_2 under squared error loss function becomes when $a_0 = b_0 = a_1 = b_1 = a_2 = b_2 = 0$ becomes

$$\widehat{\lambda}_{1B}(\alpha) = \frac{n(r_1 + r_4)}{T(\alpha)(n_1 + r_1)}$$

$$\widehat{\lambda}_{2B}(\alpha) = \frac{n(r_2 + r_5)}{T(\alpha)(n_1 + r_1)}$$

The Bayes estimate coincide with the corresponding MLEs.

Credible Set

A 100(1-γ)% credible set of λ_1 and λ_2 is

$$P((\lambda_1, \lambda_2) \in C_{\alpha,1-\gamma}(\lambda_1, \lambda_2)) = 1 - \gamma,$$

when

$$(\lambda_1, \lambda_2) \sim \pi(\lambda_1, \lambda_2|, \alpha).$$

Result

The following lemma will be useful. If

$$(X, Y) \sim \mathrm{BG}(b_0 + T(\alpha), A_0, A_1, A_2)$$

then

$$Z = X + Y \sim \mathrm{Gamma}(A_0, b_0 + T(\alpha))$$
$$V = \frac{X}{X + Y} \sim \mathrm{Beta}(A_1, A_2),$$

and Z and V are independent.

Credible Set:

A credible set $C_{\alpha,1-\gamma}(\lambda_1, \lambda_2)$ will be of the form

$$\left\{(\lambda_1, \lambda_2); \lambda_1 > 0, \lambda_2 > 0, A \leq \lambda_1 + \lambda_2 \leq B, C \leq \frac{\lambda_1}{\lambda_1 + \lambda_2} \leq D\right\}$$

where

$$P(A \leq Z \leq B) \times P(C \leq V \leq D) = 1 - \gamma$$

HPD Credible Set

It simply follows that $C_{\alpha,1-\gamma}(\lambda_1, \lambda_2)$ is a trapezoid enclosed by the following four straight lines:

$$(i) x+y = A, \;\; (ii) x+y = B, \;\; (c) x(1-D) = yD, \;\; (d) x(1-C) = yC.$$

The area of the trapezoid becomes $(B^2 - A^2)(D - C)/2$. Therefore to find the HPD credible interval we need to choose A, B, C, D, such that $(B^2 - A^2)(D - C)$ is minimum when

$$P(A \leq Z \leq B) \times P(C \leq V \leq D) = 1 - \gamma.$$

Shape Parameter is Unknown

In this case the joint posterior density of λ_1, λ_2 and α can be written as

$$\pi(\lambda_1, \lambda_2, \alpha | Data) = \pi(\lambda_1, \lambda_2 | Data, \alpha) \times \pi(\alpha | Data),$$

where

$$\pi(\alpha | Data) = k \times \pi_2(\alpha) \times \alpha^{n_1} \times$$

$$\prod t_i^{\alpha} \times \frac{1}{(b_0 + T(\alpha))^{A_0 - 1}}$$

Bayes Estimates

The Bayes estimates cannot be obtained in closed form. The following result will be useful to obtain Bayes estimates:

Result: The posterior density function of α, i.e. $l(\alpha|Data)$, is log-concave.

Therefore, it is possible to generate samples directly from $l(\alpha|Data)$, or it can be approximated very well by a gamma distribution.

Algorithm

1. Generate α from $\pi(\alpha|Data)$
2. Generate λ_1, λ_2 from $\pi(\lambda_1, \lambda_2|Data, \alpha)$
3. Repeat this process to generate $\{(\alpha_1, \lambda_{11}, \lambda_{21}), \ldots, (\alpha_M, \lambda_{1M}, \lambda_{2M})\}$
4. Once you have the generated samples from the posterior distribution, it is possible to compute the Bayes estimate or the HPD credible interval of any function of the parameter.

Credible Set

In this case a 100(1-γ)% credible set becomes $C_{1-\gamma}(\lambda_1, \lambda_2, \alpha)$, where

$$P((\lambda_1, \lambda_2, \alpha) \in C_{1-\gamma}(\lambda_1, \lambda_2, \alpha)) = 1 - \gamma,$$

where

$$(\lambda_1, \lambda_2, \alpha) \sim \pi(\lambda_1, \lambda_2, \alpha | Data).$$

Credible Set

Choose β and δ such that

$$(1-\beta)\times(1-\delta)=(1-\gamma),$$

then

$$C_{1-\gamma}(\lambda_1,\lambda-2,\alpha)=(\alpha_L,\alpha_U)\times C_{\alpha,1-\delta}(\lambda_1,\lambda_2),$$

where α_L and α_U are such that

$$\int_{\alpha_L}^{\alpha_U}\pi(\alpha|Data)d\alpha=1-\beta.$$

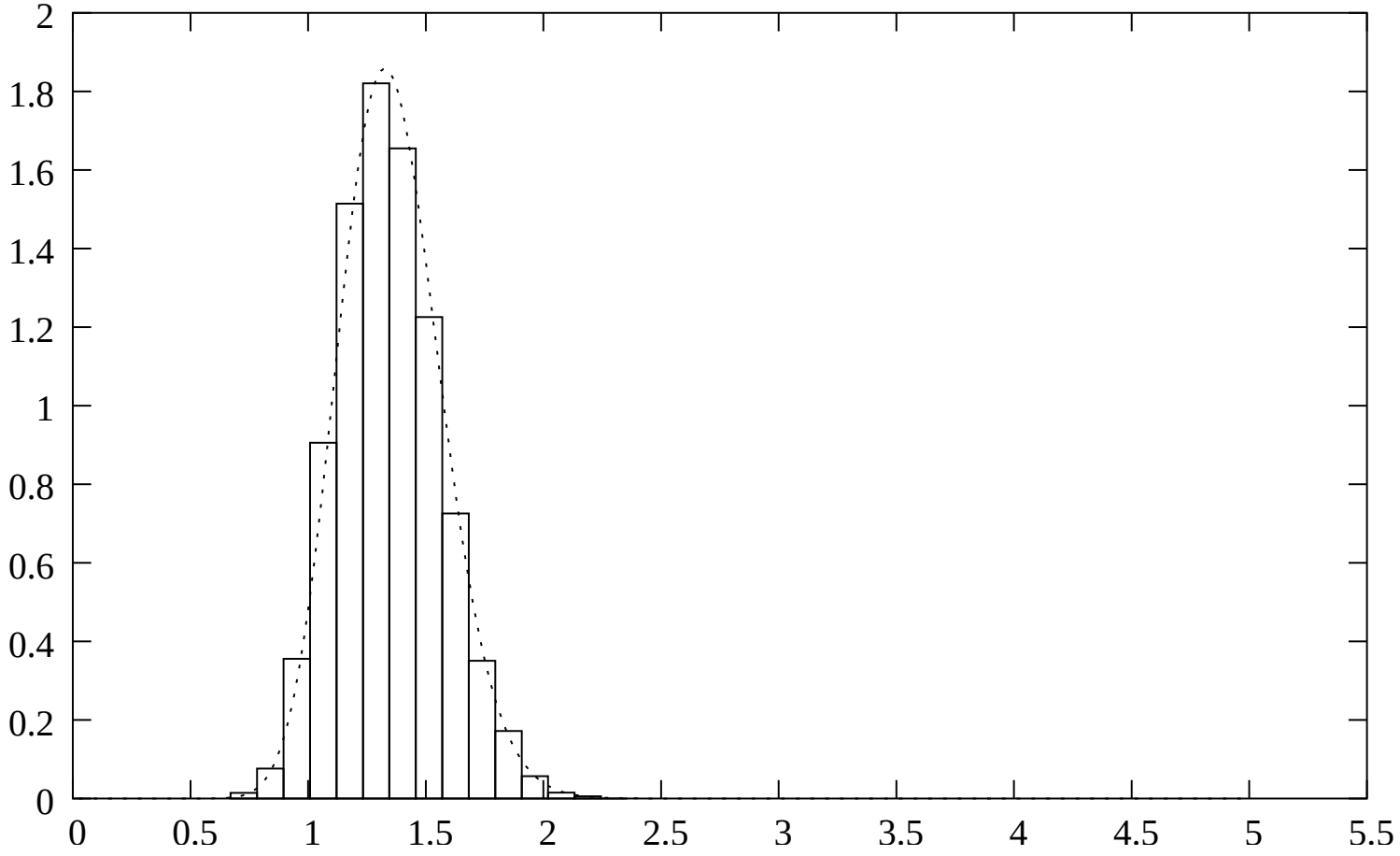
Outline

1. Competing Risks
2. Different Approaches
3. Model Formulation
4. Model and Prior Assumptions
5. Posterior Analysis
6. Bayes Estimates
7. Data Analysis
8. Analysis in Presence of Covariates

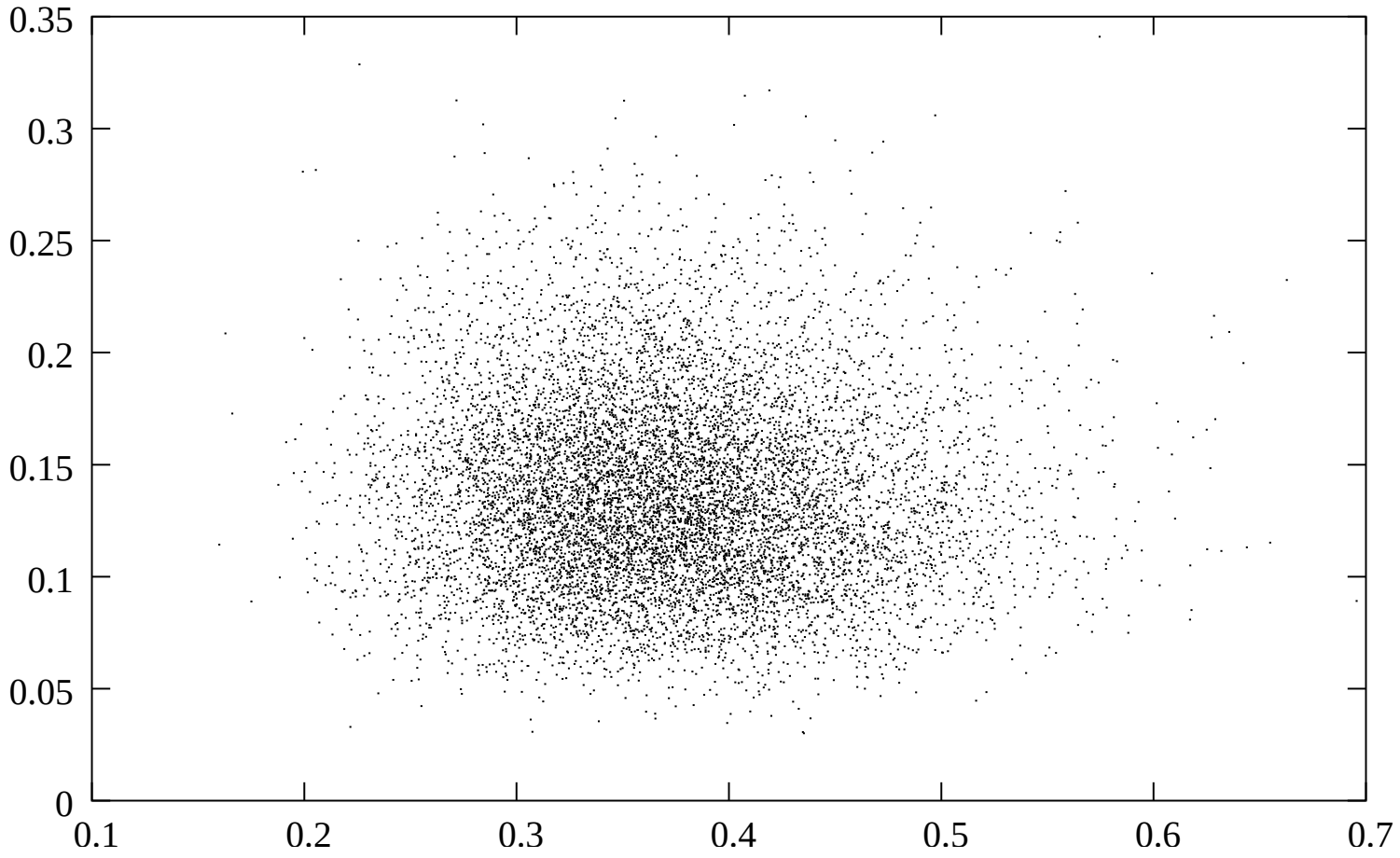
Data

1. Data set indicates 79 male stage 4 cancer patients.
2. Survival time (in weeks) indicates whether a patient is judged asymptomatic or symptomatic.
3. Approximately 35% survival times are censored.
4. Approximately 50% lack classification.

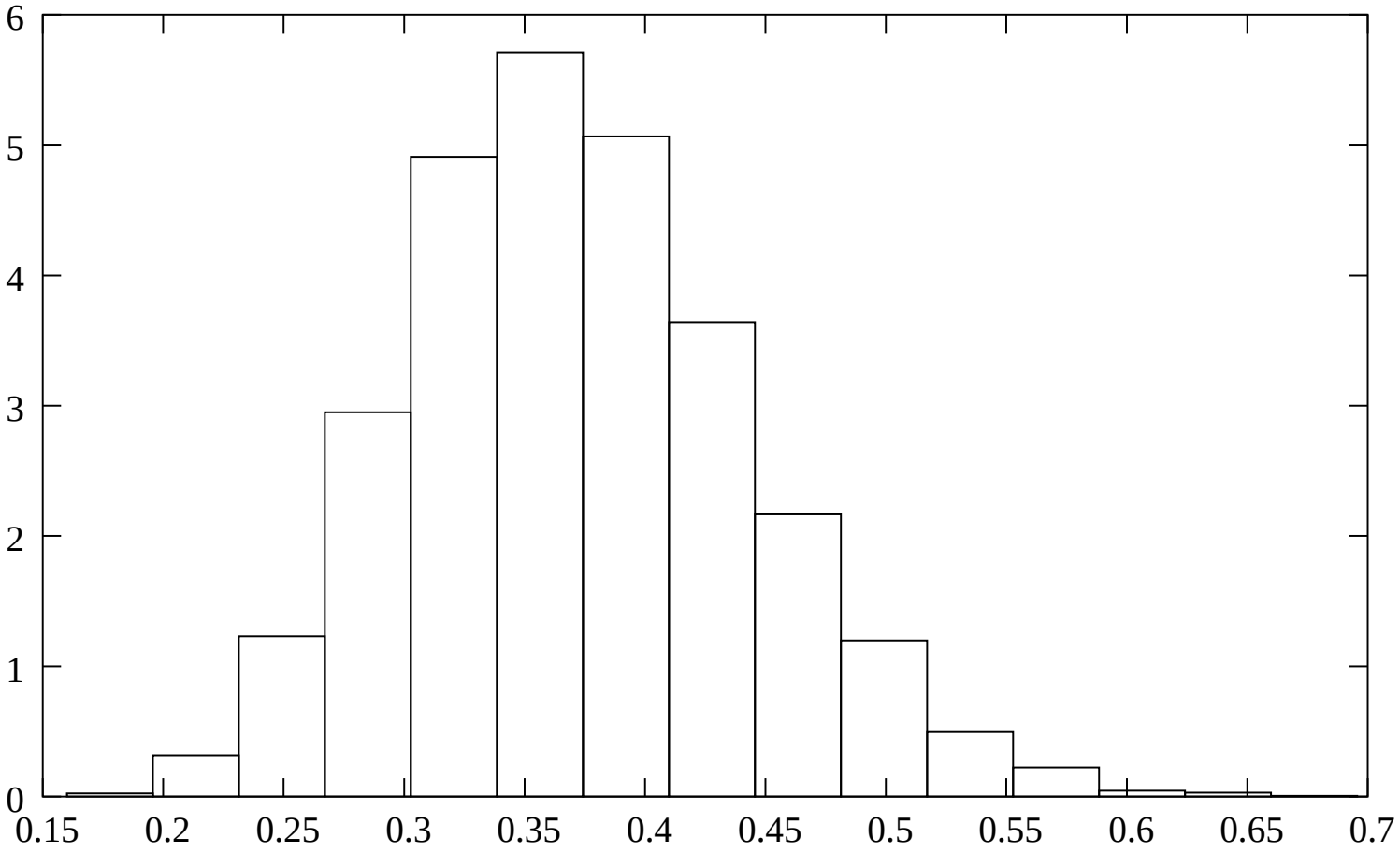
Generated samples from posterior distribution of α



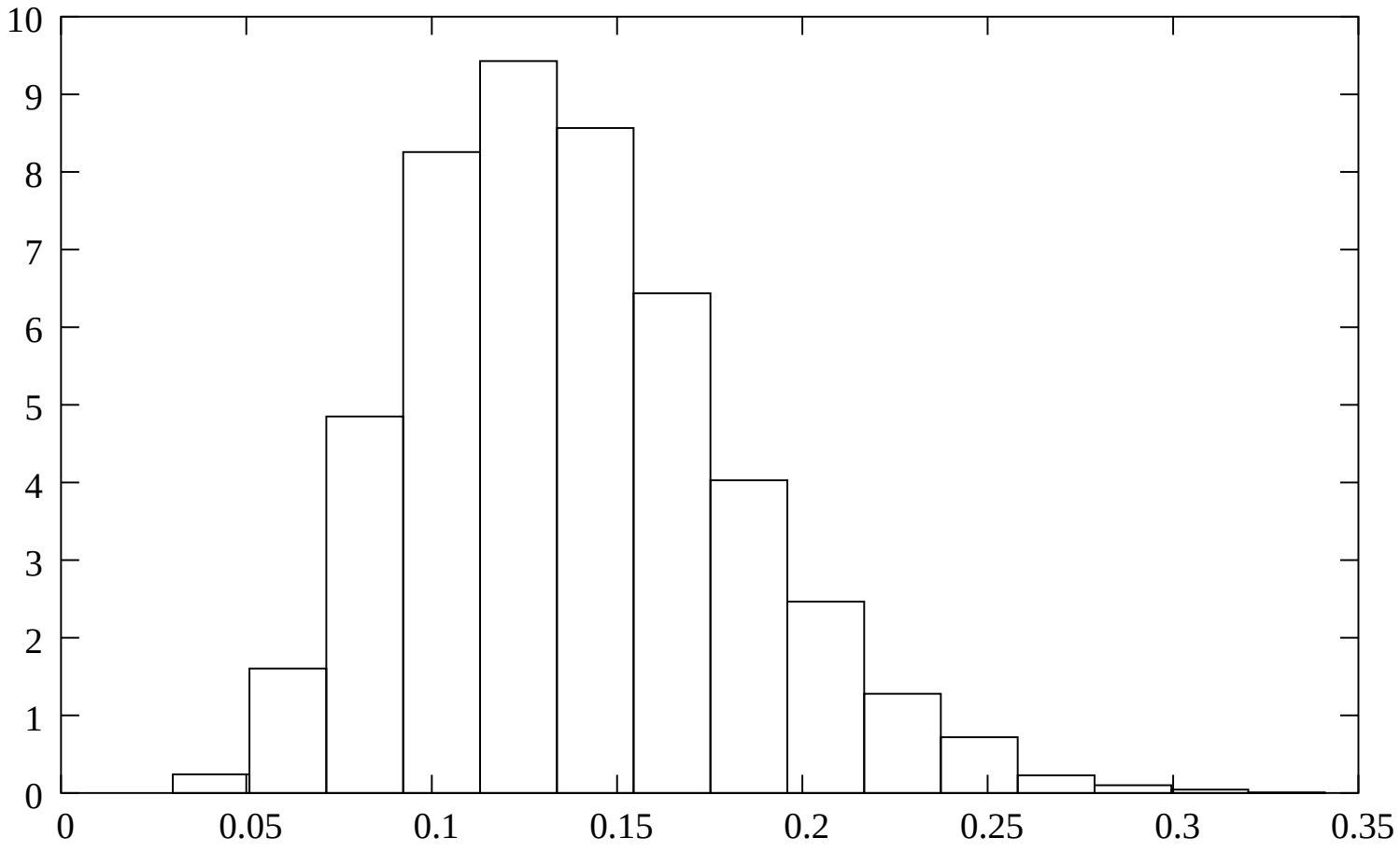
Generated samples from posterior distribution of (λ_1, λ_2)



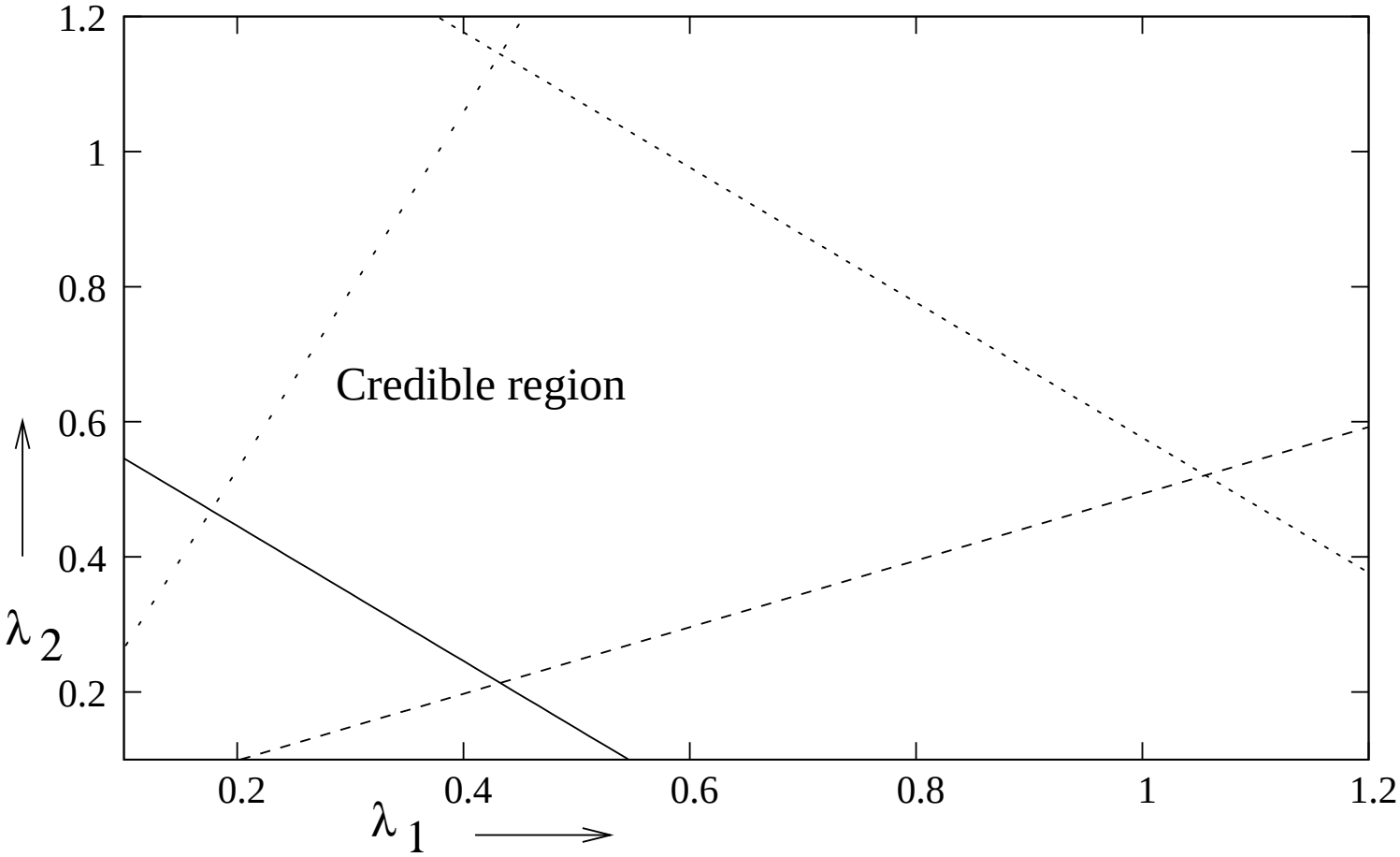
Generated samples from posterior distribution of λ_1



Generated samples from posterior distribution of λ_2



Credible Set



Data Analysis Results:

The Bayes estimates of α, λ_1 and λ_2 under squared error loss functions become

$$\widehat{\alpha} = 1.3427, \quad \widehat{\lambda}_1 = 0.3719 \quad \widehat{\lambda}_2 = 0.1377$$

The associated 95% HPD credible intervals become

$$(0.9532, 1.8021), \quad (0.2476, 0.5213), \quad (0.0664, 0.2348),$$

respectively.

Outline

1. Competing Risks
2. Different Approaches
3. Model Formulation
4. Model and Prior Assumptions
5. Posterior Analysis
6. Bayes Estimates
7. Data Analysis
8. Analysis in Presence of Covariates

Available Data & Likelihood Contribution

It is assumed that for each individual there exists a set of covariate vector x. Hence we have the following types of observations:

$$(t, 1, x), \quad (t, 2, x), \quad (t, *, x), \quad (t*, *, x), \quad (t*, 1, x), \quad (t*, 2, x).$$

Therefore, likelihood contribution from the observations from different sets are as follows:

$$f_1(t; x)S_2(t; x), \quad f_2(t; x)S_1(t; x), \quad f(t; x), \quad S(t; x)$$

$$\int_t^\infty f_1(y; x)S_2(y; x)dy, \quad \int_t^\infty f_2(y; x)S_1(y; x)dy.$$

Model and Prior Assumptions

It is assumed that the scale parameters of the latent failure time distributions depend on the covariates as follows:

$$\lambda_{1i} = \theta_1 \exp(\beta^T x_i) \quad \text{and} \quad \lambda_{2i} = \theta_2 \exp(\beta^T x_i)$$

and

$$(\theta_1, \theta_2) \sim \pi_1(\theta_1, \theta_2) = \mathrm{BG}(b_0, a_0, a_1, a_2)$$

$$\alpha \sim \pi_2(\alpha)$$

$$\beta_j \sim \pi_3(\beta_j) = N(0, \sigma_j^2)$$

Posterior Analysis

The posterior distribution takes the following form:

$$l(\theta_1, \theta_2, \alpha, \beta | Data) = k \times \mathrm{BD}(T(\alpha, \beta) + b_0, s_0, s_1, s_2) \times$$

$$g_1(\alpha | Data) \times g_2(\beta | Data) \times h(\alpha, \beta).$$

Here $g_1(\alpha|Data)$ is log-concave and $g_2(\beta|Data)$ is the product of normal distribution.
Importance sampling method can be used to compute the Bayes estimate and the associated credible interval for any function of the parameter.

Data Analysis

In this case we have a data set of 65 cancer patients. In this case the survival time in days and the corresponding cause of death, either cancer or otherwise are reported. For some patients the cause of death is missing (not known). For each patients we have used single covariate namely age. Here the observations fall into three categories, namely $(t, 1, x), (t, 2, x), (t, *, x)$. The Bayes estimates of α, θ_1, θ_2 and β are

$$2.812, \quad 0.448, \quad 0.110 \quad -0.164.$$

The associated 95% credible intervals become

$$(2.659, 3.034), \quad (0.382, 0.543), \quad (0.081, 0.128), \quad (-0.369, 0.134).$$

Thank You