Statistics and Applications {**ISSN 2452-7395 (online)**}
Volume 16, No. 1, 2018 (New Series), pp 315-337

Polya Tree Priors for the Proportional Hazards Model

Lei Huang[1] **and Malay Ghosh**[2]
[1]*Division of Biostatistics, Office of Biostatistics and Epidemiology, CBER, FDA*
[2]*Department of Statistics, University of Florida*

Received: December 18, 2017; Revised: May 29, 2018; Accepted: June 10, 2018

Abstract

The paper revisits estimation problem for the Cox proportional hazards model. We have considered the problem in a fully Bayesian nonparametric framework with Polya tree priors for the baseline survival function. The reason for choosing Polya tree priors is that it can select absolutely continuous probability measures with probability 1, given proper choice of parameters. Our finding improves the partial likelihood approach in the sense that it takes both ranks and spacings of the order statistics into account. Also, it generalizes the findings of Muliere and Walker Muliere (1997) by including covariates.

Key words: Spacings; Bayesian nonparametric method; Cox model

1 Introduction

Cox (1972) introduced the proportional hazards (PH) model by specifying the hazard rate at time t for an individual with covariate vector $\boldsymbol{x}$. The hazard rate $h(t \mid \boldsymbol{x})$ is given by

$$h(t \mid \boldsymbol{x}) = h_0(t) exp(\boldsymbol{x^T \beta}), \tag{1.1}$$

where $h_0(t)$ is the baseline hazard function. The model implies that the ratio of the hazards for two individuals is constant over time provided that the covariates stay the same over time. The proportional hazards model has been used extensively for its mathematical simplicity and straightforward interpretation.

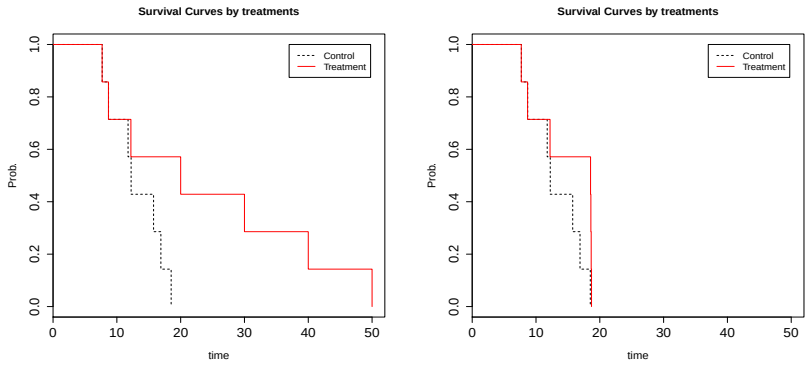
Suppose we observe censored data $(t_k, \delta_k, \boldsymbol{x}_k)$, $(k = 1, \cdots, n)$, where $t_1 < t_2 < \cdots < t_n$. Here δ_k is the censoring indicator, with $\delta_k = 0$ denoting that t_k is a right censored observation. The regression parameter vector $\boldsymbol{\beta}$ is usually estimated by maximizing the partial likelihood function $PL(\boldsymbol{\beta})$, where

$$PL(\boldsymbol{\beta}) = \prod_{k=1;\delta_k=1}^{n-1} \frac{exp(x_k^T \beta)}{\sum_{j=k+1}^{n} exp(x_j^T \beta)}. \tag{1.2}$$

Corresponding Author: Malay Ghosh
E-mail: ghoshm@ufl.edu

It should be noted that $PL(\boldsymbol{\beta})$ is a conditional likelihood function instead of a conventional likelihood function, and it depends on the ranks of event times only. This might be undesirable in some circumstances where it is believed that the spacings of order statistics should also play a role.

To illustrate the effects of spacings of order statistics, we present a simple toy example. For simplicity, we assume $\boldsymbol{x}$ is an 1-dimensional covariate, with $x_1 = 1$ for subjects in the experiment group and $x_1 = 0$ for subjects in the control group. In both scenarios in Figure (1), the black dashed lines denote the same control group ($x_1 = 0$), while the red lines are survival curves of the treatment group ($x_1 = 1$).



(a) (b)

Figure 1: Two examples with same rank order statistics

Rank order statistics in both scenarios are exactly the same but it is not the case for spacings. It is not surprising that the estimates of treatment effects based on the partial likelihood function are the same as given in Table (1). This is not quite ideal because apparently spacings play a very important role in both cases. In scenario 1(a), it seems plausible that the treatment group has a larger mean survival time than that of the control group. Hence, ideally we would like to include the spacing effect in the model and end up with a more significant treatment effect. It is the opposite for scenario 1(b). Therefore, we are proposing to take both ranks and spacings of order statistics into account, via a Bayesian nonparametric method.

Table 1: Partial Likelihood Function Estimates

	coef	**exp(coef)**	**se(coef)**	**P-value**	**lower .95**	**upper .95**
Treatment	-1.1051	0.0.3312	0.7082	0.119	0.08265	1.327

Bayesian nonparametric methods have received extensive attention recently because of their flexibility. Dirichlet process priors, introduced by Ferguson (1973), are probably the most commonly used nonparametric priors. Over the years, a large body of theory has developed for the use of such priors in survival analysis. For example, Susarla and Van Ryzin (1976) derived the Bayes estimator of the survival function under the Dirichlet process prior and Ferguson and Phadia (1979) derived the posterior distribution of the cumulative distribution function with right censored data.

However, their scope is somewhat limited due to the fact that Dirichlet process selects, with probability 1, discrete probability measures. In many areas, for example, in survival analysis, people usually deal with continuous random variables rather than discrete ones. Therefore, the need for a nonparametric prior that picks continuous probability measures could not be achieved by Dirichlet process priors.

For years, people have been seeking other approaches to accommodate continuity. For instance, Kalbfleisch (1978) defined a family of random probabilities called the Gamma process. Dykstra and Laud (1981) specified an extended Gamma process prior on the hazard rate. Hjort (1990) discussed Beta process in the context of survival analysis. We are interested in Polya tree priors, originally introduced by Fabius (1964) to provide a density estimator on $[0, 1]$ that was weakly consistent. It was later termed as a Polya tree process by Ferguson (1974). These priors can select continuous distributions with positive probability and, if necessary, even with probability 1. Lavine (1992) Lavine (1994) investigated the basic but very important properties of Polya tree priors. Sufficient conditions for these priors to assign probability 1 to the set of continuous distributions are discussed in Mauldin, Sudderth, and Williams (1992) (MSW) and Lavine (1992). Thus, it is a natural idea to employ this method to survival analysis.

More recently, people tend to use a mixture of nonparametric process priors instead of a single one due to increasing computing capability. The mixture of Dirichlet process priors, introduced by Antoniak (1974), offers a reasonable compromise between purely parametric and purely nonparametric methods, and is used extensively. Doss (1994) and Doss and Huffer (2003) discussed implementation of mixtures of Dirichlet priors for $F(t) = 1 - S(t)$ in the presence of right censored data using Gibbs sampler. Mixtures of Polya tree priors are relatively new. In particular, Hanson and Johnson (2002) have used a mixture of Polya tree priors in accelerated failure time (AFT) models. Hanson (2006) and Hanson and Yang (2007) fitted the PH model using mixture of Polya trees, and compared the PH model to the AFT model as well as the proportional odds model. Hanson and Jara (2013) fitted fully nonparametric survival models based on generalizations of the Dirichlet process mixture and the Polya tree, and compared these nonparametric models to simpler ones, like the PH model and the AFT model. In comparison to a specific Bayesian nonparametric process prior, the mixture contains a wider class of distributions and the mixture parameter smooths the prior to some extent, but it also brings in certain complexity. Approximate calculations are usually employed for mixtures of Polya tree priors in the sense that the Polya tree is truncated at a finite number of steps. Our interest lies in the case where a non-truncated Polya tree process is used. We show that the marginal likelihood function can be written in an explicit form. Muliere (1997) discussed how Polya tree priors might be used in survival analysis without covariates. Their original idea of making the Polya tree partition dependent on data greatly simplifies the problem. Huang and Ghosh (2014) used the same idea to perform a hypothesis test for Lehmann alternatives. We further extend their idea to solve estimation problems in the PH model in occurrence of covariates.

Marginal likelihoods for regression coefficients in the Cox proportional hazards model with a Bayesian nonparametric prior on the baseline survival function have been studied for years. Kalbfleisch and Prentice (1973) proved that, without ties, a marginal likelihood is well approximated by a partial likelihood with a very diffuse Gamma process prior on the baseline hazard

function. This result is restricted to the case of continuous univariate survival data and fixed time-constant covariates. Sinha, Ibrahim, and Chen (2003) generalized these results by establishing Bayesian justification of partial likelihoods as limiting cases of marginal distributions of the regression parameters under the grouped data likelihood, continuous-data likelihood with time-varying covariates and regression coefficients. A similar result is obtained in our setup. A closed form for the marginal likelihood function of the regression coefficients in the Cox proportional hazards model is derived, assuming that the baseline hazard function is drawn from a Polya tree process. Also, the marginal likelihood resembles the partial likelihood function with some natural heuristic interpretation in a limiting case.

The structure of the paper is as follows. The definition and basic properties of Polya tree priors as considered here are reviewed in Section 2. In Section 3, the marginal likelihood function of $\boldsymbol{\beta}$ is given, and some properties are discussed. In addition, a Bayesian analysis is presented. A real data analysis is given in Section 4. Finally in Section 5, we discuss the problem when ties occur and also give some pointers for future work.

2 Polya Trees

The Polya tree process is a large class of priors that includes the Dirichlet process as a special case. It provides a flexible way for Bayesian analysis of nonparametric problems. It has two parts, a partition of the sample space and a large set of non-negative parameters. Unlike the Dirichlet process, the partition plays a deterministic role in the Polya trees, and the large collection of parameters makes it possible to incorporate a wide range of variability. The following definition of the Polya tree process is taken from Lavine (1992, 1994) and KalbfGhosh and Ramamoorthileisch (2003).

Let $E = \{0, 1\}$, $E^0 = \emptyset$. Let E^m be the m-fold product $E \times E \times E \cdots \times E$ and $E^* = \bigcup_0^\infty E^m$. Define a separating binary tree of partition of Ω, $\Pi = \{\pi_m, m = 0, 1, 2, ...\}$, such that $\pi_0 = \Omega$. Also, $\pi_0, \pi_1, ...$ form a sequence of partitions such that $\bigcup_0^\infty \pi_m$ generates the measurable sets and every $B \in \pi_{m+1}$ is obtained by splitting some $B' \in \pi_m$ into two sets. Degenerate splits are permitted, i.e. some $B \in \pi_m$ can be split into $B \bigcup \emptyset$.

For each m, $\pi_m = \{B_{\vec{\epsilon_m}} : \vec{\epsilon_m} = \epsilon_1, ..., \epsilon_m \in E^m\}$ is a partition of Ω such that for all $\vec{\epsilon_m} \in E^*$, $B_{\vec{\epsilon_m},0}$, $B_{\vec{\epsilon_m},1}$ is a partition of $B_{\vec{\epsilon_m}}$. Let $A = \{a_{\vec{\epsilon_m}} : \vec{\epsilon_m} \in E^*\}$ be a set of nonnegative real numbers and $\mathfrak{y} = \{Y_{\vec{\epsilon_m}} : \vec{\epsilon_m} \in E^*\}$ be a collection of random variables. Then we say a random probability measure P on Ω have a Polya tree distribution with parameter (Π, A), written $P \sim PT(\Pi, A)$, if the following conditions hold:

1. all the random variables in $\mathfrak{y}$ with subscripts ending with 0 are independent, i.e. $Y_{\vec{\epsilon_m},0}$, for all $\vec{\epsilon_m} \in E^*$, are independent; $Y_{\vec{\epsilon_m},1} = 1 - Y_{\vec{\epsilon_m},0}$;

2. for every $\vec{\epsilon_m} \in E^*$, $Y_{\vec{\epsilon_m},0}$ has a Beta distribution with parameters $a_{\vec{\epsilon_m},0}$ and $a_{\vec{\epsilon_m},1}$;

3. for every $m = 1, 2, ...$ and every $\vec{\epsilon_m} \in E^*$,

$$P(B_{\epsilon_1,...,\epsilon_m}) = (\prod_{j=1;\epsilon_j=0}^{m} Y_{\epsilon_1,...,\epsilon_j}) \prod_{j=1;\epsilon_j=1}^{m} (1 - Y_{\epsilon_1,...,\epsilon_{j-1},0}) = \prod_{j=1}^{m} Y_{\epsilon_1,...,\epsilon_j} \tag{2.1}$$

Here the form of $P(B_{\epsilon_1,...,\epsilon_m})$ differs from what is given in Lavine (1992) by re-arranging $Y_{\vec{\epsilon}_m}$ and defining $Y_{\vec{\epsilon}_m,1} = 1 - Y_{\vec{\epsilon}_m,0}$. This provides a compact expression for $P(B_{\epsilon_1,...,\epsilon_m})$, noting that $Y_{\vec{\epsilon}_m,0}$ and $Y_{\vec{\epsilon}_m,1}$ are not independent.

Some properties of Polya tree processes are listed below to conclude this section. For more properties, see Lavine (1992) and KalbfGhosh and Ramamoorthileisch (2003).

Remark: Basic properties of the Polya tree process:

1. The Polya trees are conjugate. If P has a Polya tree distribution, and $X \mid P \sim P$, then $P \mid X$ has a Polya tree distribution. The posterior distribution is updated in the following manner: for every $\vec{\epsilon_m}$ such that $X \in B_{\vec{\epsilon_m}}$, add 1 to $a_{\vec{\epsilon_m}}$.

2. Some Polya trees assign probability 1 to the set of continuous distributions. A broadly used sufficient condition for this can be found in Theorem 3.3.7 in KalbfGhosh and Ramamoorthileisch (2003). For example, take $a_{\vec{\epsilon_m}} = cm^2$, where c is a positive real number.

If we have a Polya tree with partitions $\{B_{\vec{\epsilon_m}} : \vec{\epsilon_m} \in E^*\}$ and parameters A, the predictive density at $x \in B_{\vec{\epsilon_m}}$ is given by

$$\begin{aligned} f(x) &= \lim_{m\to+\infty} \frac{Pr(B_{\vec{\epsilon_m}})}{\lambda(B_{\vec{\epsilon_m}})} \\ &= \lim_{m\to+\infty} \frac{\prod_{i=1}^{m} \frac{a_{\epsilon_1,...,\epsilon_j}}{a_{\epsilon_1,...,\epsilon_{j-1},0} + a_{\epsilon_1,...,\epsilon_{j-1},1}}}{\lambda(B_{\vec{\epsilon_m}})} \end{aligned} \tag{2.2}$$

where $\lambda(\cdot)$ is the Lebesgue measure.

A Polya tree can be constructed by centering at an arbitrary distribution. This is important because when we use Polya tree priors, we usually have a guess of the underlying distribution. Thus by proper construction, we can make the expectation of the Polya tree coincide with our guess. There are two ways to do this. Suppose $\Omega = \mathbb{R}$, say we want a Polya tree to center at a pre-specified probability measure G, which is referred to as the baseline measure of the Polya tree. Let the partition be such that the elements of π_m are taken as the intervals $[G^{-1}(k/2^m), G^{-1}((k+1)/2^m))$ for $k = 0, 1, ..., 2^m - 1$, with the obvious interpretation for $G^{-1}(0)$ and $G^{-1}(1)$. We will refer to this as Method 1. The other approach is to make the partition data-dependent, as mentioned in Muliere (1997). This would make the calculation for the posterior with censored data much simpler. Suppose we specify a number of points $x_1 < ... < x_n$ as end points, and let $B_1 =$

$[x_1, +\infty)$, $B_{11} = [x_2, +\infty)$,...,$B_{\underbrace{1,...,1}_{n}} = [x_n, +\infty)$. In addition we need the parameters $a_{\vec{\epsilon_m}}$ to satisfy

$$\frac{a_{\epsilon_1,...,\epsilon_{j-1},0}}{a_{\epsilon_1,...,\epsilon_{j-1},1}} = \frac{G(B_{\epsilon_1,...,\epsilon_{j-1},0})}{G(B_{\epsilon_1,...,\epsilon_{j-1},1})}$$

and that $a_{\vec{\epsilon_m}}$ increases quickly enough to ensure the continuity property. Here and later, any other unspecified subintervals are generated by splitting their parent intervals into two equal parts with respect to the G measure. For example, if $B_0 = (0, x_1)$, the sets B_{00} and B_{01} are such that $G(B_{00}) = G(B_{01}) = G(B_0)/2$. This we refer as Method 2. We will take $a_{\vec{\epsilon_m}} = cm^2$ to ensure continuity of the priors.

Method 2 yields some extra benefits. For example, it assigns probability 1 to a set of continuous probability measures. The probability density function exists and is defined as in (2.2). The partition and parameters in Method 2 differ from Method 1 only by finitely many terms. Thus if we calculate the limit in (2.2) for any $x \in \mathbb{R}^+$, the limit exists and is finite. We carried out the calculations based on the partitions described in Method 2. As we will see in later sections, the results only depend on finitely many parameters, $a_{\underbrace{1,...,1}_{k}}$ and $a_{\underbrace{1,...,1}_{k-1},0}$ for $k = 1, \cdots, n$. The calculations will go through as long as Polya trees select continuous distributions with probability 1 and $a_{\epsilon_1,...,\epsilon_m}$ grows to infinity as $m \to +\infty$.

3 Cox Proportional Hazards Model Under Polya Tree Process Prior

3.1 Marginal Likelihood Function for β

We consider the case with right censoring in survival data analysis. The data are of the form $(t_k, \delta_k, \boldsymbol{x}_k)$, where $\boldsymbol{x_k}$ is the set of covariates associated with the kth subject, and δ_k is the censoring indicator, with $\delta_k = 1$ being the case that the kth observation is an event.

Assuming Cox proportional hazards model as in (1.1), it follows that the survival function is given by

$$S(t \mid \boldsymbol{x}) = [S_0(t)]^{exp(\boldsymbol{x}^T\boldsymbol{\beta})}.$$

Kalbfleisch and Prentice (1973) derived the marginal likelihood functions of the Cox model with and without ties. Later, Kalbfleisch (1978) used Gamma process (GP) for the baseline and derived the marginal likelihood function of the coefficients. In this section we present results analogous to ones obtained by Kalbfleisch's work. Instead of putting a Gamma process prior for the baseline, we use a Polya tree prior.

The Polya tree prior appears to be more flexible than Gamma process prior. To see this, suppose

$$P_1 \sim GP(\gamma(t), \beta) \quad and \quad P_2 \sim PT(\Pi, A).$$

Suppose $t_1 < t_2 < t_3$ are such that

$$\gamma(t_2) - \gamma(t_1) = \gamma(t_3) - \gamma(t_2)$$

Then by the construction of the GP, we have

$$E[P_1((t_1, t_2))] = E[P_1((t_2, t_3))] \quad and \quad Var[P_1((t_1, t_2))] = Var[P_1((t_2, t_3))].$$

Thus it is clear that the variance is determined homogeneously at some level.

However, this is not the same for Polya trees. By choosing proper parameters, it is possible to make

$$E[P_2((t_1, t_2))] = E[P_2((t_2, t_3))] \quad and \quad Var[P_2((t_1, t_2))] \neq Var[P_2((t_2, t_3))].$$

This is useful because in some cases, the researcher might not be sure about the distribution over a certain interval, and wants to let the distribution have a relatively large variance in that interval. Polya tree can deal with this situation.

For example, suppose a horrible earthquake took place at t_2 and aftershocks kept coming during t_2 to t_3. This might bring in a lot of censored data within that period. One might want to let the variance of the baseline distribution be bigger during (t_2, t_3) than that during (t_1, t_2). A Polya tree prior can accommodate situations like this.

Throughout this section, ties are not considered. We assume that the censoring occurs infinitesimally before an event happens. Namely, if we observe a censored datum t, the contribution of this datum to likelihood is $Pr([t, +\infty))$. In case one needs to deal with the opposite situation, one should let the partition be $B_{\underbrace{1,...,1}_{i}} = (X_i, +\infty)$, for $i = 1, \cdots, n$, and redo the computations.

Now suppose that the baseline distribution $F_0 = 1 - S_0$ is drawn from a Polya tree process. Our interest is to estimate $\boldsymbol{\beta}$. For the time being, we are not considering this problem in a fully Bayesian framework. That is, no prior is given for $\boldsymbol{\beta}$ and thus finding the posterior of $\boldsymbol{\beta}$ is not our primary goal. The reason is that the structure of Polya tree is complicated, that even if we use the simplest prior for $\boldsymbol{\beta}$, the posterior is not available analytically. Hence, we confine our interest in the marginal likelihood of $\boldsymbol{\beta}$ and seek Maximum Likelihood Estimator (MLE) throughout this subsection, which is also the posterior mode under a uniform prior for $\boldsymbol{\beta}$.

The following theorem gives the exact form of the marginal likelihood function of $\boldsymbol{\beta}$. The proof of the theorem is provided as supplemental material.

Theorem 1: Assume the model (1.1) and no ties occur in the data. Suppose the baseline distribution is distributed as a Polya tree $PT(\Pi, A)$. Then the exact likelihood function of coefficients

$\boldsymbol{\beta}$ is given by $Lik = \lim_{m\to+\infty} E[L_m]$, where

$$
\begin{aligned}
E[L_m] = &\prod_{k=1}^{n} \{\frac{1}{\lambda(B_{\underbrace{1,...,1}_{k}\underbrace{0,...,0}_{m-k}})}\}^{\delta_k} \prod_{k=1}^{n-1} \{a_{\underbrace{1,...,1}_{k},0}\}^{\delta_k} \{\frac{a_{\underbrace{1,...,1}_{n},0}}{a_{\underbrace{1,...,1}_{n},0} + a_{\underbrace{1,...,1}_{n},1}}\}^{\delta_n} \\
&\cdot \prod_{k=1}^{n} \{\frac{a_{\underbrace{1,...,1}_{k},00}}{a_{\underbrace{1,...,1}_{k},00} + a_{\underbrace{1,...,1}_{k},01}} \frac{a_{\underbrace{1,...,1}_{k},000}}{a_{\underbrace{1,...,1}_{k},000} + a_{\underbrace{1,...,1}_{k},001}} \cdots \frac{a_{\underbrace{1,...,1}_{k},\underbrace{0,...,0}_{m-k}}}{a_{\underbrace{1,...,1}_{k},\underbrace{0,...,0}_{m-k}} + a_{\underbrace{1,...,1}_{k},\underbrace{0,...,0}_{m-k-1},1}}\}^{\delta_k} \\
&\cdot \{\prod_{k=1}^{n} exp(x_k^T\beta)^{\delta_k}\} \cdot \frac{\Gamma(a_0+a_1)\Gamma(a_1+\sum_{j=1}^{n} exp(x_j^T\beta))}{\Gamma(a_0+a_1+\sum_{j=1}^{n} exp(x_j^T\beta))\Gamma(a_1)} \\
&\cdot \{\prod_{k=1}^{n-1} \frac{\Gamma(a_{\underbrace{1,...,1}_{k},0} + a_{\underbrace{1,...,1}_{k},1})\Gamma(a_{\underbrace{1,...,1}_{k},1} + \sum_{j=k+1}^{n} exp(x_j^T\beta))}{\Gamma(a_{\underbrace{1,...,1}_{k},0} + a_{\underbrace{1,...,1}_{k},1} + \sum_{j=k+1}^{n} exp(x_j^T\beta) + \delta_k)\Gamma(a_{\underbrace{1,...,1}_{k},1})}\}. \qquad (3.1)
\end{aligned}
$$

Note that the terms involving $\boldsymbol{\beta}$ are independent of m. To get the MLE of $\boldsymbol{\beta}$, we only need to maximize the following function, $L(\beta)$, with respect to $\boldsymbol{\beta}$, where

$$
\begin{aligned}
L(\beta) = &\{\prod_{k=1}^{n} exp(x_k^T\beta)^{\delta_k}\} \cdot \frac{\Gamma(a_1+\sum_{j=1}^{n} exp(x_j^T\beta))}{\Gamma(a_0+a_1+\sum_{j=1}^{n} exp(x_j^T\beta))} \\
&\cdot \{\prod_{k=1}^{n-1} \frac{\Gamma(a_{\underbrace{1,...,1}_{k},1} + \sum_{j=k+1}^{n} exp(x_j^T\beta))}{\Gamma(a_{\underbrace{1,...,1}_{k},0} + a_{\underbrace{1,...,1}_{k},1} + \sum_{j=k+1}^{n} exp(x_j^T\beta) + \delta_k)}\}. \qquad (3.2)
\end{aligned}
$$

It is worth pointing out that the exact likelihood function of $\boldsymbol{\beta}$ is just $L(\boldsymbol{\beta})$ multiplied by a positive constant, i.e.$Lik(\boldsymbol{\beta}) = c_0 L(\boldsymbol{\beta})$. If we re-parameterize as follows,

$$
\begin{cases}
\sigma_k &= a_{\underbrace{1,...,1}_{k}} + a_{\underbrace{1,...,1}_{k-1},0} \\
r_k &= \frac{G([v_k,+\infty))}{G([v_{k-1},+\infty))} = \frac{a_{\underbrace{1,...,1}_{k}}}{a_{\underbrace{1,...,1}_{k-1},0} + a_{\underbrace{1,...,1}_{k}}}
\end{cases}, \qquad (3.3)
$$

then $L(\beta)$ reduces to

$$
\begin{aligned}
L(\beta) = &\{\prod_{k=1}^{n} exp(x_k^T\beta)^{\delta_k}\} \cdot \frac{\Gamma(\sigma_1 r_1+\sum_{j=1}^{n} exp(x_j^T\beta))}{\Gamma(\sigma_1+\sum_{j=1}^{n} exp(x_j^T\beta))} \\
&\cdot \{\prod_{k=1}^{n-1} \frac{\Gamma(\sigma_{k+1} r_{k+1} + \sum_{j=k+1}^{n} exp(x_j^T\beta))}{\Gamma(\sigma_{k+1} + \sum_{j=k+1}^{n} exp(x_j^T\beta) + \delta_k)}\}. \qquad (3.4)
\end{aligned}
$$

Moreover, when $\sigma_1 = ... = \sigma_n = 0$, $L(\beta)$ reduces to

$$\begin{aligned} L(\beta) =& \{\prod_{k=1}^{n} exp(x_k^T\beta)^{\delta_k}\}\{\prod_{k=1}^{n-1} \frac{\Gamma(\sum_{j=k+1}^{n} exp(x_j^T\beta))}{\Gamma(\sum_{j=k+1}^{n} exp(x_j^T\beta) + \delta_k)}\} \\ =& \prod_{k=1;\delta_k=1}^{n-1} \frac{exp(x_k^T\beta)}{\sum_{j=k+1}^{n} exp(x_j^T\beta)} \end{aligned}$$

which resembles the partial likelihood function for the Cox proportional hazards model. Recall that the partial likelihood function is

$$\begin{aligned} PL(\beta) =& \prod_{k=1;\delta_k=1}^{n} \frac{exp(x_k^T\beta)}{\sum_{j=k}^{n} exp(x_j^T\beta)} \\ =& \prod_{k=1;\delta_k=1}^{n-1} \frac{exp(x_k^T\beta)}{\sum_{j=k}^{n} exp(x_j^T\beta)} \end{aligned}$$

They only differ in that $exp(x_k^T\beta)$ is not included in the denominator of each term in the product. If one thinks of the motivation of each term in the partial likelihood function heuristically as a conditional probability given by

$$\begin{aligned} Pr(\textit{the particular individual dies at } t_k \mid \textit{one death at } t_k) =& \frac{h(t_k \mid x_k)}{\sum_{j=k}^{n} h(t_j \mid x_j)} \\ =& \frac{exp(x_k^T\beta)}{\sum_{j=k}^{n} exp(x_j^T\beta)}, \end{aligned}$$

then the special case of $L(\beta)$ is based on conditional odds instead of conditional probabilities for each term, namely,

$$\begin{aligned} Odds(\textit{the particular individual dies at } t_k \mid \textit{one death at } t_k) =& \frac{\frac{h(t_k|x_k)}{\sum_{j=k}^{n} h(t_j|x_j)}}{1 - \frac{h(t_k|x_k)}{\sum_{j=k}^{n} h(t_j|x_j)}} \\ =& \frac{exp(x_k^T\beta)}{\sum_{j=k+1}^{n} exp(x_j^T\beta)} \end{aligned}$$

3.2 Effect of Spacings

In this subsection, we investigate the effect of spacings of order statistics on the MLE. For simplicity, we consider only one-dimensional covariate in this subsection. Without loss of generality, assume that the covariate is non-negative. We consider individually for the component terms in (3.4). For $k = 1, \cdots, n-1$, let

$$h(r_{k+1}) = log(\Gamma(\sigma_{k+1}r_{k+1} + \sum_{j=k+1}^{n} exp(x_j\beta))) - log(\Gamma(\sigma_{k+1} + \sum_{j=k+1}^{n} exp(x_j\beta) + \delta_k)),$$

and

$$h(r_{k+1})_\beta = [\psi(\sigma_{k+1}r_{k+1} + \sum_{j=k+1}^{n} exp(x_j\beta)) - \psi(\sigma_{k+1} + \sum_{j=k+1}^{n} exp(x_j\beta) + \delta_k)](\sum_{j=k+1}^{n} exp(x_j\beta)x_j),$$

where $\psi(\cdot)$ is the Digamma function. $h(r_{k+1})_\beta$ is increasing in r_{k+1} because $\psi(\cdot)$ is strictly increasing in $(0, +\infty)$ and $\sum_{j=k+1}^{n} exp(x_j\beta)x_j > 0$ unless all covariates are trivially 0. This is also true for $h(r_1)$ where $h(r_1)$ is defined in the obvious way. Thus the MLE is obtained by solving

$$h(\beta) = \sum_{k=0}^{n-1} h(r_{k+1})_\beta = 0. \tag{3.5}$$

Note that $h(\beta)$ is an increasing function of r_k, $k = 1, \cdots, n$. Hence larger r_k's result in larger $h(\beta)$, and accordingly result in larger β's, i.e. larger treatment effects, as shown in Figure (2).

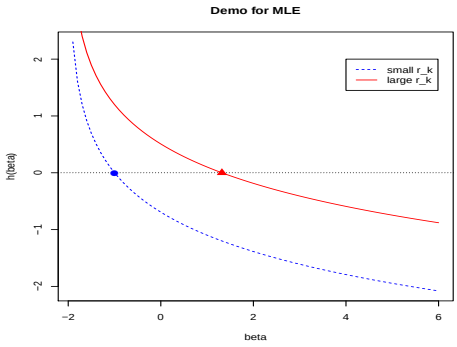


Figure 2: Effects of spacings on MLE

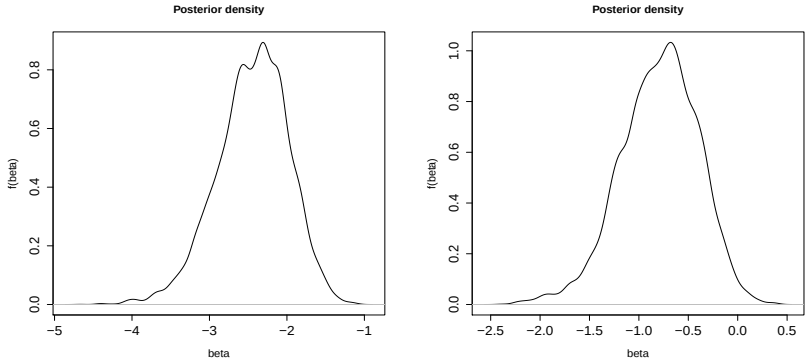
3.3 Full Bayesian Analysis

A fully Bayesian setup is available in this framework and is discussed in this sub-section. Assume that a generic prior $\pi(\boldsymbol{\beta})$ is assigned to β. It is straightforward that the posterior is given by

$$\begin{aligned}\pi(\boldsymbol{\beta} \mid data) &\propto Lik(\boldsymbol{\beta})\pi(\boldsymbol{\beta}) \\ &\propto L(\boldsymbol{\beta})\pi(\boldsymbol{\beta}).\end{aligned}$$

Since an explicit expression for $L(\boldsymbol{\beta})$ is available, it is straightforward to apply standard Metropolis-Hastings algorithm to simulate the posterior distribution of $\boldsymbol{\beta}$.

Here we revisit the toy example (Figure (1)) proposed in Section 1. A relatively non-informative N(0,100) prior is assigned to $\boldsymbol{\beta}$. Figure (3) shows the kernel density estimates of the posteriors of $\boldsymbol{\beta}$ using MCMC algorithm.



(a) Mean=-2.44, 95% HPD=(-3.38,-1.57) (b) Mean=-0.79, 95% HPD=(-1.53,-0.003)

Figure 3: Kernel Density for Posteriors of β

Compared to the partial likelihood estimate, $\beta_{PL} = -1.1051$, in Table (1), the posterior mean of β in scenario 1(a) is evaluated at -2.44, showing a sign of much larger treatment effect; the posterior mean of β in scenario 1(b) is evaluated at -0.79, showing a smaller treatment effect. Another interesting result is that the highest posterior density (HPD) interval for scenario 1(a) is $(-3.38, -1.57)$, which is clearly far away from 0, identifying a statistically significant treatment effect. In contrast, the HPD interval for scenario 1(b) is $(-1.53, -0.003)$, which barely excludes 0, making the significance of treatment effect less convincing.

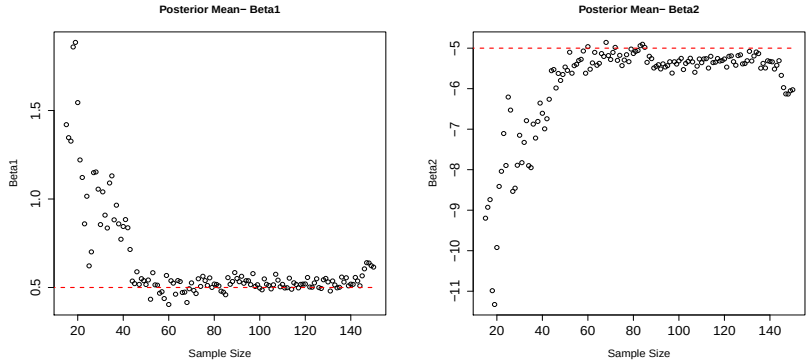
3.4 Simulation Study Showing Consistency of the Posterior Mean

To study asymptotic properties of the estimation, a series of simulation studies were carried out. A Weibull (3,12) is used as the baseline survival distribution for simulation . Observations are censored by an independent Exp(50) distribution. Assume that there are 2 covariates. The true regression coefficients are set to be 0.5 and -5 respectively. We randomly generated censored sample with sizes from 15 to 150. Throughout all calculations for posteriors, a non-informative prior, N (0,100), was assigned to the two coefficients independently. For each sample size, a MCMC algorithm is used to compute the posterior mean. The posterior means are plotted in Figure (4).

As one can see in the simulations, the posterior means of the regression coefficients converge to the true values quickly and stay stable ever since.

4 Real Data Analysis

We take the ovarian cancer dataset in the Survival package of R software as a real data analysis example. The data set was originally reported by Edmunson, Fleming and Decker (1979), and was



(a) (b)

Figure 4: Posterior Means of β's as Sample Size Increases

later re-analyzed in a number of situations, such as Collett (2003). The study included $n = 26$ patients with advanced ovarian carcinoma (stages IIIB and IV). Treatment of patients using either cyclophosphamide alone (1 g/m2) or cyclophosphamide (500 mg/m2) plus adriamycin (40 mg/m2) by injection every 3 weeks each produced partial improvement in approximately one third of the patients.

Table 2: **Ovarian Cancer**

Treatment	Survival Time (Days)
Treatment 1	59, 115, 156, 268, 329, 431, 448+, 477+, 638, 803+, 855+, 1040+, 1106+
Treatment 2	353, 365, 377+, 421+, 464, 475, 563, 744+, 769+, 770+, 1129+, 1206+, 1227+

As an illustration, we consider the Bayesian parametric model described by Dellaportas and Smith (1993). The likelihood under the Weibull model is given by

$$L(\boldsymbol{\beta}, \rho \mid data) = \{\prod_{j=1}^{n} \rho t_j^{\rho-1} exp(\boldsymbol{x_j}\boldsymbol{\beta})\}\{\prod_{j=1}^{n+m} exp[-t_j^{\rho} exp(\boldsymbol{x_j}\boldsymbol{\beta})]\}. \tag{4.1}$$

We assume that $\beta_1, \beta_2, \beta_3 \sim N(0, 100)$ independently, where $\beta_1, \beta_2, \beta_3$ are regression coefficients correponding to treatment, age and residual disease respectively. Due to the fact that for ECOG score, 1 is better than 2, we assign a $lognormal\ (0, 100)$ prior to β_4. Through Gibbs sampling in Winbugs, the posterior densities are given in Figure (5).

Now we proceed with our proposed method. A relatively noninformative prior is used. The same priors for $\boldsymbol{\beta}$ are used. Standard Metropolis within Gibbs sampler algorithm can be applied and the kernel density estimates of resulting posterior distributions are displayed in Figure (6).

A simple comparison of the results shows that our method leads to a treatment effect with larger magnitude. Furthermore, our method successfully identifies a positive age effect. This is reasonable as one would expect that older people are more likely to have larger hazards.

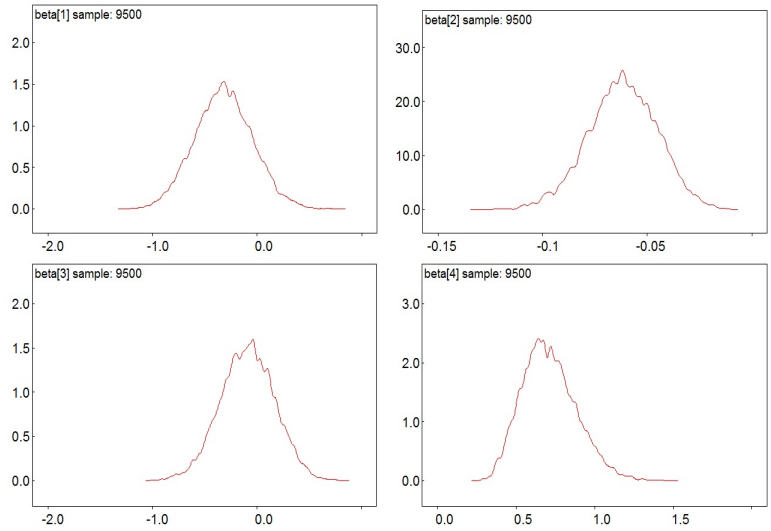


Figure 5: Kernel Density for Posteriors of β's (Dellaportas and Smith's method Dellaportas and Smith (1993))

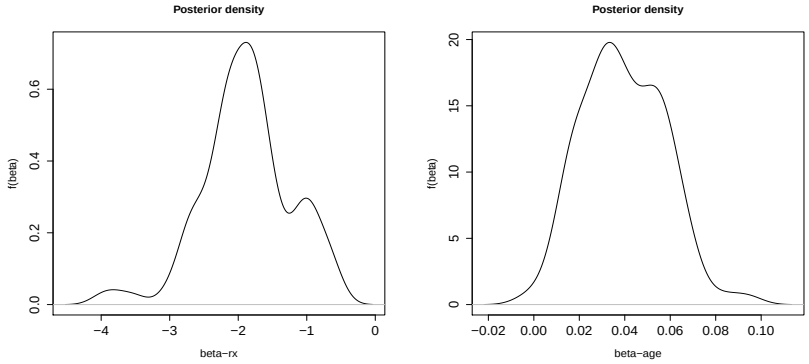
Hennerfeind (2006) introduced a method using B-spline to model the baseline hazard as an alternative nonparametric approach. Hennerfeind, Brezger, and Fahrmeir (2006) further included time-varying covariates as well as a spatial component for geographical effects in this nonparametric model. We adopt the idea of B-spline and present the results for comparison. In this model, a cubic spline with 2 knots was generated. The coefficients of the Spline basis functions and the regression coefficients for age, treatment, residual disease and ECOG performance score are estimated by maximizing the full likelihood function, with the constraint that the coefficients for the Spline are nonnegative. It appears the Spline model overestimates treatment effects and it does not capture the trend that higher ECOG scores should be associated with higher risks (hazards). The results are presented in Table (3) and the estimate of the baseline hazard function is shown in Figure (7).

Table 3: **Ovarian Cancer - Spline Model**

	Age	**Residual Disease**	**Treatment**	**ECOG**
Estimates	0.0430	-0.0113	-3.3262	-1.0490

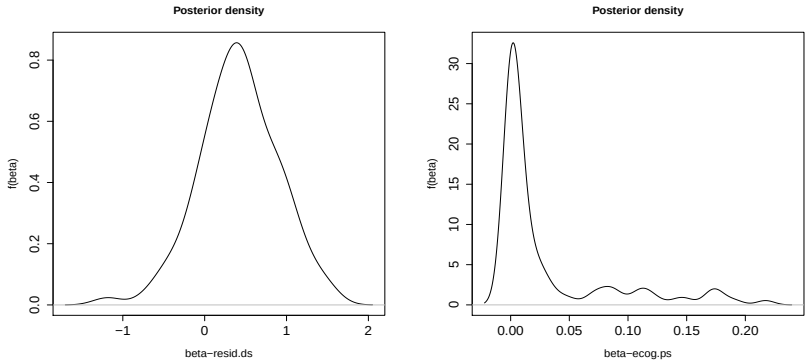
5 Discussion

So far our approach only works in case without ties. However, in reality, there might be situations where ties occur. For example, public emergencies might cause multiple events or censoring to occur at the same time.We divide ties into two categories, namely ties of censored data and ties of event times. For the first case when multiple observations are censored at the same time, our calculations are still valid. However, this is not the case for the second scenario.



(a) Mean=-1.88, 95% HPD=(-2.908,-0.578)

(b) Mean=0.039, 95% HPD=(0.0099,-0.071)



(c) Mean=0.416, 95% HPD=(-0.514,1.379)

(d) Mean=0.031, 95% HPD=(0,0.170)

Figure 6: Kernel Density for Posteriors of β's

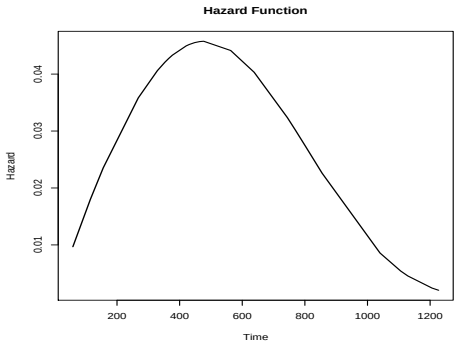


Figure 7: Baseline hazard function - Spline Model

The problem occurs for two reasons. First, the limit (2.2) may not exist or be infinity for some $x \in \mathbb{R}^+$ because the density does not necessarily exist when the underlying distribution is

not absolutely continuous. Second, when multiple events are observed at the same time, we know that the underlying distribution must be non-continuous. In this case, to keep the prior reasonable, we should not assign parameters such that Polya tree gives probability 1 to a set of continuous distributions. Instead, appropriate parameters should be assigned to the Polya tree such that it gives positive probability either to a set of discrete distributions, or to a set of partly discrete and partly continuous distributions. Dirichlet process appears to be a natural candidate in this case. However, it turns out to be computationally prohibitive, since the likelihood function depends not only on the number of tied observations, but also on the location on $\mathbb{R}^+$ of the occurrences of ties. The problem becomes intractable as sample size increases, since the number of possible combinations of ties grows exponentially.

One way to overcome the hurdles is to learn from how people deal with ties in the partial likelihood function. Suppose observed data set $(t_k, \delta_k, \boldsymbol{x}_k)$, $k = 1, 2, \cdots, n$ (without loss of generality, assuming data are sorted by t_k) has tied event times $t_{k_0} = t_{k_0+1} = t$. Assume that the underlying distribution of times is continuous. In this case, we observe ties because of measurement error. In reality, $t_{k_0} \neq t_{k_0+1}$, but they are so close to each other such that we do not see the difference. Just by looking at the data, there could be two equally likely possibilities, i.e. $t_{k_0} > t_{k_0+1}$ and $t_{k_0} < t_{k_0+1}$ by a small margin. Let ϵ be a very small positive real number. Then we can calculate approximate $L(\boldsymbol{\beta})$ by letting both $t_{k_0} = t + \epsilon$, $t_{k_0+1} = t$ and $t_{k_0+1} = t + \epsilon$, $t_{k_0} = t$. Then the overall $L(\boldsymbol{\beta})$ is given by the average of resulting two $L(\boldsymbol{\beta})$'s since both possibilities are equally likely. The continuity of $L(\boldsymbol{\beta})$ on r_{k_0} guarantees that approximate $L(\boldsymbol{\beta})$ is close to its true value as long as ϵ is small.

References

Muliere P and Walker S. (1997). A Bayesian non-parametric approach to survival using Polya trees. *Scandinavian Journal of Statistics*, **24**, 331–340.

Cox D. R. (1972) Regression models and life-tables. *Journal of the Royal Statistical Society, Series B.* **34**, 187–220.

Ferguson T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, **1**, 209–230.

Susarla V and Van Ryzin J. (1976). Nonparametric Bayesian estimation of survival curves from incomplete observations. *Journal of the American Statistical Association*, **71**, 897–902.

Ferguson T. S. and Phadia E. G. (1979). Bayesian nonparametric estimation based on censored data. *The Annals of Statistics*, **7**, 163–186.

Kalbfleisch J .D. ((1978). Non-parametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society, Series B*, **40**, 214–221.

Dykstra R. L. and Laud P. (1981). A Bayesian nonparametric approach to reliability. *The Annals of Statistics*, **9**, 356–367.

Hjort N. L. (1990). Nonparametric Bayes estimators based on Beta processes in models for life history data. *The Annals of Statistics*, **18**, 1259–1294.

Fabius J. (1964). Asymptotic behavior of Bayes' Estimates. *Annals of Mathematical Statistics*, **35**, 846–856.

Ferguson T. S. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics*, **2**, 615–629.

Lavine M. (1992). Some aspects of Polya tree distributions for statistical modelling. *The Annals of Statistics*, **20**, 1222–1235.

Lavine M. (1994). More aspects of Polya tree distributions for statistical modelling. *The Annals of Statistics*, **22**, 1161–1176.

Mauldin R. D., Sudderth W. D. and Williams S. C. (1992). Polya trees and random distributions. *The Annals of Statistics*, **20**, 1203–1221.

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

Doss H. (1994). Bayesian nonparametric estimation for incomplete data via successive substitution sampling. *The Annals of Statistics*, **22**, 1763–1786.

Doss H and Huffer F. (2003). Monte Carlo methods for Bayesian analysis of survival data using mixtures of Dirichlet priors. *Journal of Computational and Graphical Statistics*, **12**, 282–307.

Hanson T. E. and Johnson WO.(2002). Modeling regression error with a mixture of Polya trees. *Journal of American Statistical Association*, **97**, 1020–1033.

Hanson T. E. (2006). Inference for mixtures of finite Polya tree models. *Journal of American Statistical Association*, **101**, 1548–1565.

Hanson T. E. and Yang M. (2007). Bayesian semiparametric proportional odds models. *Biometrics*, **63**, 88–95.

Hanson T. E and Jara A. (2013). *Surviving fully Bayesian nonparametric regression models. Bayesian theory and applications.* DOI:10.1093/acprof:oso/9780199695607.003.0030

Huang L and Ghosh M. (2014). Two-sample hypothesis testing under Lehmann alternatives and Polya tree priors. *Statistica Sinica*, **24**, 1717–1733.

Kalbfleisch J. D and Prentice R. L. (1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika.* **60**, 267–278.

Sinha D, Ibrahim J. G. and Chen M. H. (2003). A Bayesian justification of Cox's partial likelihood. *Biometrika*, **90**, 629–641.

Ghosh J. K. and Ramamoorthi R. V. (2003). *Bayesian Nonparametrics.* Springer, New York

Edmunson J. H., Fleming T. R., Decker G. D. et al. (1979). Different chemotherapeutic sensitivities and host factors affecting prognosis in advanced ovarian carcinoma versus minimal disease residual. *Cancer Treatment Reports.* **63**, 241–247.

Collett, D.(2003). *Modelling survival data in medical research.* Chapman and Hall, Boca Raton

Dellaportas P and Smith A. F. (1993). Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling. *The Applied Statistics*, **42**,443–459.

Hennerfeind A. (2006). Bayesian nonparametric regression for survival and event history data [Dissertation]. Munchen (Germany): Ludwig-Maximilians-Univerisity

Hennerfeind A, Brezger A and Fahrmeir L. (2006). Geoadditive survival models. *Journal of American Statistical Association*, **101**, 1065–1075.

Appendix

The proof of Theorem 1 is provided.

Proof. Take $m > n$, such that at level m, $t_1, ..., t_n$ are separated in different intervals. For $t \in [0, +\infty)$, let $\vec{\epsilon_m}(t) = \epsilon_1, ..., \epsilon_m$ such that $t \in B_{\epsilon_1,...,\epsilon_m}$. In addition, with appropriate parameters, P is continuous with probability 1. Thus without loss of generality, assume $t_1 < ... < t_n$. And write $\vec{\epsilon_m}^{\,i} = \vec{\epsilon_m}(t_i) = \epsilon_1^i, ..., \epsilon_m^i$.

When there are no covariates, $t_1, ..., t_n \mid P \sim P = 1 - \bar{P}$. At level m of the tree, given P, the conditional independence yields the joint pdf of $t_1, ..., t_n$ is

$$\begin{aligned} f_m(t_1, ..., t_n \mid P) &= \frac{\prod_{i=1}^n Pr(B_{\vec{\epsilon_m}^{\,i}} | P)}{\prod_{i=1}^n \lambda(B_{\vec{\epsilon_m}^{\,i}})} \\ &= \frac{\prod_{i=1}^n \prod_{\epsilon_j^i=0} Y_{\epsilon_1^i,...,\epsilon_j^i} \prod_{\epsilon_j=1} (1 - Y_{\epsilon_1^i,...,\epsilon_j^i})}{\prod_{i=1}^n \lambda(B_{\vec{\epsilon_m}^{\,i}})} \end{aligned} \tag{5.1}$$

The exact marginal joint pdf is given by letting $m \to +\infty$ and then taking the expectation. By repeated use of the Theorem 2 in Lavine Lavine (1992), the existence and finiteness of the limit are guaranteed. We denote this limit by f. By dominated convergence theorem, the order of expectation and limit can be interchanged.

Now with the occurrence of covariates x_k, write

$$\alpha_k = exp(x_k^T \beta)$$

for $k = 1, ..., n$.

We will calculate the conditional likelihood function of (t_k, d_k), given P, by considering censored and uncensored data separately. Let us begin with the uncensored situation. Without loss of

generality, assume that all data are uncensored, i.e., $d_k = 1$, for $k = 1, \cdots, n$. The conditional joint likelihood function is not trivial in this case. Take any $t \in B_{\epsilon_1,...,\epsilon_m}$. We need to find $P_{\alpha_k}(B_{\vec{\epsilon_m}})$, where P_{α_k} denotes the probability measure induced by $P_{\alpha_k} = 1 - \bar{P}^{\alpha_k}$. To do this, we have to sum all the probabilities of the intervals to the right of $B_{\vec{\epsilon_m}}$, plus probability of $B_{\vec{\epsilon_m}}$, and then raise to power α_k. This quantity is denoted by $P_{\alpha_k}(B_{\vec{\epsilon_m}}+ \mid P)$. At mth level of the tree, given P, a simple expression is provided by,

$$P_{\alpha_k}(B_{\vec{\epsilon_m}}+ \mid P) = \{\sum_{n=1}^{m} Pr(B_{\epsilon_1,...,\epsilon_{n-1},1} \mid P)\delta_0(\epsilon_n) + Pr(B_{\vec{\epsilon_m}} \mid P)\}^{\alpha_k}.$$

Therefore,

$$\begin{aligned} &P_{\alpha_k}(B_{\vec{\epsilon_m}} \mid P) \\ &= P_{\alpha_k}(B_{\vec{\epsilon_m}}+ \mid P) - P'((B_{\vec{\epsilon_m}}+) - B_{\vec{\epsilon_m}} \mid P) \\ &= \{\sum_{n=1}^{m} Pr(B_{\epsilon_1,...,\epsilon_{n-1},1} \mid P)\delta_0(\epsilon_n) + Pr(B_{\vec{\epsilon_m}} \mid P)\}^{\alpha_k} - \{\sum_{n=1}^{m} Pr(B_{\epsilon_1,...,\epsilon_{n-1},1} \mid P)\delta_0(\epsilon_n)\}^{\alpha_k} \end{aligned} \tag{5.2}$$

where the "-" sign in the probability means exclusion. Now using the second order Taylor Expansion for function $h(t) = t^\alpha$,

$$h(t+\Delta) - h(t) = \alpha t^{\alpha-1}\Delta + \alpha(\alpha-1)(t+\theta)^{\alpha-2}\Delta^2$$

where $\theta \in (0, \Delta)$. It follows that

$$\begin{aligned} P_{\alpha_k}(B_{\vec{\epsilon_m}} \mid P) = \quad & \alpha_k \quad Pr(B_{\vec{\epsilon_m}} \mid P)\{\sum_{n=1}^{m} Pr(B_{\epsilon_1,...,\epsilon_{n-1},1} \mid P)\delta_0(\epsilon_n)\}^{\alpha_k - 1} \\ & + \quad \alpha_k(\alpha_k - 1)Pr(B_{\vec{\epsilon_m}} \mid P)^2\{\sum_{n=1}^{m} Pr(B_{\epsilon_1,...,\epsilon_{n-1},1} \mid P)\delta_0(\epsilon_n) + \theta\}^{\alpha_k - 2} \end{aligned}$$

where $\theta \in (0, Pr(B_{\vec{\epsilon_m}} \mid P))$. For simplicity, write

$$W_m(t) = \sum_{n=1}^{m} Pr(B_{\epsilon_1,...,\epsilon_{n-1},1} \mid P)\delta_0(\epsilon_n)$$

W_m depends on t because $\epsilon_1, ..., \epsilon_m$ depend on t.

Now we are in place to calculate the conditional joint pdf of $t_1, \cdots, t_n$ at mth level of the Polya tree, which is denoted by $\bar{f}_m(t_1, ..., t_n \mid P)$.

$$\begin{aligned} &\bar{f}_m(t_1, ..., t_n \mid P) \\ &= \frac{\prod_{i=1}^{n} P_{\alpha_i}(B_{\vec{\epsilon_m}^{\,i}} \mid P)}{\prod_{i=1}^{n} \lambda(B_{\vec{\epsilon_m}^{\,i}})} \\ &= \frac{\prod_{i=1}^{n} Pr(B_{\vec{\epsilon_m}^{\,i}} \mid P) \prod_{i=1}^{n} \{\alpha_i W_m(t_i)^{\alpha_i - 1} + \alpha_i(\alpha_i - 1)[W_m(t_i) + \theta_i]^{\alpha_i - 2} Pr(B_{\vec{\epsilon_m}^{\,i}} \mid P)\}}{\prod_{i=1}^{n} \lambda(B_{\vec{\epsilon_m}^{\,i}})} \\ &= \frac{NUM}{\prod_{i=1}^{n} \lambda(B_{\vec{\epsilon_m}^{\,i}})} \end{aligned}$$

where NUM is the numerator of the fraction. The exact marginal joint pdf is found by letting $m \to +\infty$ and then taking expectation with regard to Polya tree process.

We claim that

$$\lim_{m\to+\infty} E[\frac{NUM}{\prod_{i=1}^n \lambda(B_{\vec{\epsilon_m}^{\,i}})}] = \lim_{m\to+\infty} E[\frac{\prod_{i=1}^n Pr(B_{\vec{\epsilon_m}^{\,i}} \mid P)\prod_{i=1}^n \{\alpha_i W_m(t_i)^{\alpha_i-1}\}}{\prod_{i=1}^n \lambda(B_{\vec{\epsilon_m}^{\,i}})}] \tag{5.3}$$

Indeed, if we write the products in the numerator as summation, we have

$$\begin{aligned}&\prod_{i=1}^n [\alpha_i W_m(t_i)^{\alpha_i-1} + \alpha_i(\alpha_i-1)\{W_m(t_i)+\theta_i\}^{\alpha_i-2} Pr(B_{\vec{\epsilon_m}^{\,i}} \mid P)]\\ =& \sum_{S\subseteq\Omega}[\prod_{j\in S}\alpha_i W_m(t_j)^{\alpha_i-1}\prod_{k\in S^c}\alpha_i(\alpha_i-1)\{W_m(t_k)+\theta_k\}^{\alpha_i-2}Pr(B_{\vec{\epsilon_m}^{\,k}} \mid P)]\end{aligned} \tag{5.4}$$

where $\Omega = \{1, ..., n\}$, and the summation is taken for all (proper and improper) subsets $S \subseteq \Omega$.

However, if $\mid S^c \mid \geq 1$, i.e. there exists $k_0 \in S^c$, then

$$\begin{aligned}&\mid [\prod_{j\in S}\alpha_i W_m(t_j)^{\alpha_i-1}\prod_{k\in S^c}\alpha_i(\alpha_i-1)\{W_m(t_k)+\theta_k\}^{\alpha_i-2}Pr(B_{\vec{\epsilon_m}^{\,k}} \mid P)] \mid .\\ &\leq \mid \alpha^n(\alpha-1)^{|S^c|}Pr(B_{\vec{\epsilon_m}^{\,k_0}} \mid P) \mid\end{aligned}$$

where $\alpha = \max_{i=1,\cdots,n} \alpha_i$. The above inequality uses the fact that

$$0 \leq W_m(t) \leq W_m(t)+\theta \leq W_m(t)+Pr(B_{\vec{\epsilon_m}^{\,k}} \mid P) \leq \sum_{n=1}^m Pr(B_{\epsilon_1,\ldots,\epsilon_n} \mid P) = Pr([0,+\infty) \mid P) = 1.$$

Hence, the expectation of the corresponding term in summation satisfies

$$\begin{aligned}&E[\mid \frac{\prod_{i=1}^n Pr(B_{\vec{\epsilon_m}^{\,i}} \mid P)[\prod_{j\in S}\alpha_i W_m(t_j)^{\alpha_i-1}\prod_{k\in S^c}\alpha_i(\alpha_i-1)\{W_m(t_k)+\theta_k\}^{\alpha_i-2}Pr(B_{\vec{\epsilon_m}^{\,k}} \mid P)]}{\prod_{i=1}^n \lambda(B_{\vec{\epsilon_m}^{\,i}})} \mid]\\ &\leq E[\mid \frac{\prod_{i=1}^n Pr(B_{\vec{\epsilon_m}^{\,i}} \mid P)\alpha^n(\alpha-1)^{|S^c|}Pr(B_{\vec{\epsilon_m}^{\,k_0}} \mid P)}{\prod_{i=1}^n \lambda(B_{\vec{\epsilon_m}^{\,i}})} \mid]\\ &= const.E[\frac{\{\prod_{i\neq k_0} Pr(B_{\vec{\epsilon_m}^{\,i}} \mid P)\}Pr(B_{\vec{\epsilon_m}^{\,k_0}} \mid P)^2}{\prod_{i=1}^n \lambda(B_{\vec{\epsilon_m}^{\,i}})}]\end{aligned} \tag{5.5}$$

Comparing (5.5) to (5.1), it follows that

$$\begin{aligned}E\ &[\frac{\{\prod_{i\neq k_0} Pr(B_{\vec{\epsilon_m}^{\,i}} \mid P)\}Pr(B_{\vec{\epsilon_m}^{\,k_0}} \mid P)^2}{\prod_{i=1}^n \lambda(B_{\vec{\epsilon_m}^{\,i}})}]\\ =\ & E[f_m(x_1, ..., x_n \mid P)]\prod_{j=1}^m \frac{a_{\epsilon_1^{k_0},\ldots,\epsilon_j^{k_0}} + n_{\epsilon_1^{k_0},\ldots,\epsilon_{j-1}^{k_0}} + 1}{a_{\epsilon_1^{k_0},\ldots,\epsilon_{j-1}^{k_0},0} + a_{\epsilon_1^{k_0},\ldots,\epsilon_{j-1}^{k_0},1} + n_{\epsilon_1^{k_0},\ldots,\epsilon_{j-1}^{k_0}} + 1}\end{aligned}$$

where $n_{\epsilon_1^{k_0},...,\epsilon_{j-1}^{k_0}} = \sharp\{j : X_j \in B_{\epsilon_1^{k_0},...,\epsilon_{j-1}^{k_0}}\}$.

When $m > n$, we specify the parameters as

$$a_{\epsilon_1,...,\epsilon'_{m-1}0} = a_{\epsilon_1,...,\epsilon_{j-1},1} = m^2$$

which implies that when m is large,

$$\frac{a_{\epsilon_1^{k_0},...,\epsilon_j^{k_0}} + n_{\epsilon_1^{k_0},...,\epsilon_{j-1}^{k_0}} + 1}{a_{\epsilon_1^{k_0},...,\epsilon_{j-1}^{k_0},0} + a_{\epsilon_1^{k_0},...,\epsilon_{j-1}^{k_0},1} + n_{\epsilon_1^{k_0},...,\epsilon_{j-1}^{k_0}} + 1} \to \frac{1}{2},$$

and $E[f_m(x_1, ..., x_n \mid P)]$ is finite. Therefore,

$$E[\frac{\{\prod_{i\neq k_0} Pr(B_{\epsilon_m^{\vec{}}\,i} \mid P)\} Pr(B_{\epsilon_m^{\vec{}}\,k_0} \mid P)^2}{\prod_{i=1}^n \lambda(B_{\epsilon_m^{\vec{}}\,i})}] \to 0 \quad as \quad m \to +\infty.$$

So all the terms with $\mid S^c \mid \geq 1$ in (5.4) eventually tend to 0. The only term left is when $\mid S^c \mid = 0$, i.e. $S = \Omega$, which completes the proof to the claim.

As for the censored data, their contribution to the likelihood function is simpler. The contribution of the $(t_k, d_k = 0)$ at mth level of Polya tree is just

$$P_{\alpha_k}(B_{\vec{\epsilon_m}} + \mid P) = \{\sum_{n=1}^{m} Pr(B_{\epsilon_1,...,\epsilon_{n-1},1} \mid P)\delta_0(\epsilon_n) + Pr(B_{\vec{\epsilon_m}} \mid P)\}^{\alpha_k} = W_m(t_k)^{\alpha_k} \tag{5.6}$$

Before we combine the results of (5.3) and (5.6) together, let us figure out what $\vec{\epsilon_m}(t_i)$ is. By the mechanism of partition, for $m > n$, and $i = 1, ..., n$, t_i is an end point at level i of the tree. Before the ith level, t_i lies in the right subinterval every time the current interval splits into two; after ith level, t_i would always be in the left subinterval generated by splitting the current interval that contains t_i. Thus,

$$\vec{\epsilon_m}(t_i) = \underbrace{1, ..., 1}_{i}, \underbrace{0, ..., 0}_{m-i}$$

Clearly,

$$Pr(B_{\epsilon_m^{\vec{}}\,i} \mid P) = Y_1 Y_{11} ... Y_{\underbrace{1, ..., 1}_{i}} Y_{\underbrace{1, ..., 1}_{i},0} ... Y_{\underbrace{1, ..., 1}_{i},\underbrace{0, ..., 0}_{m-i}}.$$

And by definition of $W_m(t)$,

$$W_m(t_1) = Y_1,\ W_m(t_2) = Y_1 Y_{11},\ \cdots,\ W_m(t_n) = Y_1 Y_{11} ... Y_{\underbrace{1, ..., 1}_{n}}. \tag{5.7}$$

These lead to

$$\begin{aligned}\prod_{i=1}^{n} Pr(B_{\epsilon_m^{\vec{}}\,i} \mid P) =& \{Y_1 Y_{10} Y_{100} ... Y_{1,\underbrace{0, ..., 0}_{n-1}}\} \{Y_1 Y_{11} Y_{110} Y_{1100} ... Y_{11,\underbrace{0, ..., 0}_{n-2}}\} \cdots \\ & \{Y_1 Y_{11} ... Y_{\underbrace{1, ..., 1}_{n}} Y_{\underbrace{1, ..., 1}_{n},0} Y_{\underbrace{1, ..., 1}_{n},00} ... Y_{\underbrace{1, ..., 1}_{n},\underbrace{0, ..., 0}_{m-n}}\}\end{aligned} \tag{5.8}$$

Recall that

$$Y_{\epsilon_1,...,\epsilon_j,1} = 1 - Y_{\epsilon_1,...,\epsilon_j,0} \tag{5.9}$$

for all $j = 1, ..., n$ and all $(\epsilon_1, ..., \epsilon_j)$.

Combining the results from (5.3) , (5.6), (5.7) and (5.8) yields that the contribution, at level m, of $(\alpha_k, t_k, \delta_k)$ to the likelihood is asymptotically equal to

$$\begin{aligned} L_{k,m} &= \frac{Pr(B_{\vec{\epsilon_m}\,k} \mid P) \cdot \alpha_k \cdot W_m(t_k)^{\alpha_k - 1}}{\lambda(B_{\vec{\epsilon_m}\,k})} \\ &= \frac{Y_1 Y_{11} ... \underbrace{Y_{1,...,1}}_{k} \underbrace{Y_{1,...,1,0}}_{k} ... \underbrace{Y_{1,...,1,0,...,0}}_{k\quad m-k} \cdot \alpha_k \cdot (Y_1 Y_{11} ... \underbrace{Y_{1,...,1}}_{k})^{\alpha_k - 1}}{\lambda(B_{\underbrace{1,...,1}_{k}\underbrace{0,...,0}_{m-k}})} \\ &= \frac{\underbrace{Y_{1,...,1,0}}_{k} ... \underbrace{Y_{1,...,1,0,...,0}}_{k\quad m-k} \cdot \alpha_k \cdot (Y_1 Y_{11} ... \underbrace{Y_{1,...,1}}_{k})^{\alpha_k}}{\lambda(B_{\underbrace{1,...,1}_{k}\underbrace{0,...,0}_{m-k}})} \end{aligned} \tag{5.10}$$

When $\delta = 0$, t_k is a censored time. Thus the contribution to the likelihood, given P, is

$$L_{k,m} = W_m(X_k)^{\alpha_k} = \{Y_1 Y_{11} ... \underbrace{Y_{1,...,1}}_{k}\}^{\alpha_k}$$

Combining the previous two equations, the contribution of $(\alpha_k, t_k, \delta_k)$ to the likelihood is

$$L_{k,m} = (Y_1 Y_{11} ... \underbrace{Y_{1,...,1}}_{k})^{\alpha_k} \{ \frac{\alpha_k \underbrace{Y_{1,...,1,0}}_{k} ... \underbrace{Y_{1,...,1,0,...,0}}_{k\quad m-k}}{\lambda(B_{\underbrace{1,...,1}_{k}\underbrace{0,...,0}_{m-k}})} \}^{\delta_k}$$

Thus the true likelihood function has the form

$$L = E[\lim_{m \to +\infty} \prod_{k=1}^{n} L_{k,m}]$$

Again, By Lavine Lavine (1992), the density exists and is finite. By dominated convergence theorem, we can interchange the order of the expectation and the limit.

Define

$$
\begin{aligned}
L_m =& \prod_{k=1}^{n} L_{k.m} \\
=& Y_1^{\sum_{i=1}^n \alpha_i} Y_{11}^{\sum_{i=2}^n \alpha_i} \cdots \underbrace{Y_{1,\ldots,1}^{\sum_{i=k}^n \alpha_i}}_{k} \cdots \underbrace{Y_{1,\ldots,1}^{\alpha_n}}_{n} \cdot \prod_{j=1}^{n} \{ \frac{\underbrace{Y_{1,\ldots,1,0}}_{j} \cdots \underbrace{Y_{1,\ldots,1,0,\ldots,0}}_{i \quad m-i}}{\lambda(B_{\underbrace{1,\ldots,1}_{k}\underbrace{0,\ldots,0}_{m-k}})} \}^{\delta_j} \\
=& [\prod_{k=1}^{n} \frac{\alpha_k^{\delta_k}}{\lambda(B_{\underbrace{1,\ldots,1}_{k}\underbrace{0,\ldots,0}_{m-k}})^{\delta_k}}] \\
& \cdot (1-Y_0)^{\sum_{i=1}^n \alpha_i} (1-Y_{10})^{\sum_{i=2}^n \alpha_i} Y_{11}^{\delta_1} \cdots (1-\underbrace{Y_{1,\ldots,1,0}}_{k-1})^{\sum_{i=k}^n \alpha_i} \underbrace{Y_{1,\ldots,1,0}^{\delta_{k-1}}}_{k-1} \cdots (1-\underbrace{Y_{1,\ldots,1,0}}_{n-1})^{\alpha_n} \underbrace{Y_{1,\ldots,1,0}^{\delta_{n-1}}}_{n-1} \\
& \cdot \{Y_{100} Y_{1000} \cdots Y_{1,\underbrace{0,\ldots,0}_{m-1}}\}^{\delta_1} \\
& \cdot \{Y_{1100} Y_{11000} \cdots Y_{11,\underbrace{0,\ldots,0}_{m-2}}\}^{\delta_2} \\
& \cdots \\
& \cdot \{\underbrace{Y_{1,\ldots,1,0}}_{n} \underbrace{Y_{1,\ldots,1,00}}_{n} \underbrace{Y_{1,\ldots,1,000}}_{n} \cdots \underbrace{Y_{1,\ldots,1,0,\ldots,0}}_{n \quad m-n}\}^{\delta_n}
\end{aligned}
$$

With all $Y's$ appeared in last equation being independently beta distributed, taking expectation

$$
\begin{aligned}
E[L_m] =& \prod_{k=1}^{n} \{ \frac{1}{\lambda(B_{\underbrace{1,\ldots,1}_{k}\underbrace{0,\ldots,0}_{m-k}})} \}^{\delta_k} \prod_{k=1}^{n-1} \{ \underbrace{a_{1,\ldots,1,0}}_{k} \}^{\delta_k} \{ \frac{\underbrace{a_{1,\ldots,1,0}}_{n}}{\underbrace{a_{1,\ldots,1,0}}_{n} + \underbrace{a_{1,\ldots,1,1}}_{n}} \}^{\delta_n} \\
& \cdot \prod_{k=1}^{n} \{ \frac{\underbrace{a_{1,\ldots,1,00}}_{k}}{\underbrace{a_{1,\ldots,1,00}}_{k} + \underbrace{a_{1,\ldots,1,01}}_{k}} \frac{\underbrace{a_{1,\ldots,1,000}}_{k}}{\underbrace{a_{1,\ldots,1,000}}_{k} + \underbrace{a_{1,\ldots,1,001}}_{k}} \cdots \frac{\underbrace{a_{1,\ldots,1,0,\ldots,0}}_{k \quad m-k}}{\underbrace{a_{1,\ldots,1,0,\ldots,0}}_{k \quad m-k} + \underbrace{a_{1,\ldots,1,0,\ldots,0,1}}_{k \quad m-k-1}} \}^{\delta_k} \\
& \cdot \{ \prod_{k=1}^{n} \alpha_k^{\delta_k} \} \cdot \frac{\Gamma(a_0+a_1)\Gamma(a_1+\sum_{j=1}^n \alpha_j)}{\Gamma(a_0+a_1+\sum_{j=1}^n \alpha_j)\Gamma(a_1)} \\
& \cdot \{ \prod_{k=1}^{n-1} \frac{\Gamma(\underbrace{a_{1,\ldots,1,0}}_{k} + \underbrace{a_{1,\ldots,1,1}}_{k})\Gamma(\underbrace{a_{1,\ldots,1,1}}_{k} + \sum_{j=k+1}^n \alpha_j)}{\Gamma(\underbrace{a_{1,\ldots,1,0}}_{k} + \underbrace{a_{1,\ldots,1,1}}_{k} + \sum_{j=k+1}^n \alpha_j + \delta_k)\Gamma(\underbrace{a_{1,\ldots,1,1}}_{k})} \}
\end{aligned}
$$

Then the true marginal likelihood of $\alpha's$ ($\beta's$) is given by letting $m \to +\infty$. Generally it is not likely to get a close form of this likelihood function because of the difficulty in evaluating limits. However, even though (3.1) looks complex, we do not need to evaluate it if our goal is to get the

MLE of $\beta's$. Note that $\beta's$ are involved only through $\alpha's$. And terms containing $\alpha's$ are

$$L(\vec{\alpha}) = \{\prod_{k=1}^{n} \alpha_k^{\delta_k}\} \cdot \frac{\Gamma(a_0 + a_1)\Gamma(a_1 + \sum_{j=1}^{n} \alpha_j)}{\Gamma(a_0 + a_1 + \sum_{j=1}^{n} \alpha_j)\Gamma(a_1)} \cdot \{\prod_{k=1}^{n-1} \frac{\Gamma(a_{\underbrace{1,...,1}_{k},0} + a_{\underbrace{1,...,1}_{k},1})\Gamma(a_{\underbrace{1,...,1}_{k},1} + \sum_{j=k+1}^{n} \alpha_j)}{\Gamma(a_{\underbrace{1,...,1}_{k},0} + a_{\underbrace{1,...,1}_{k},1} + \sum_{j=k+1}^{n} \alpha_j + \delta_k)\Gamma(a_{\underbrace{1,...,1}_{k},1})}\},$$

where $\alpha_k = exp(x_k^T \beta)$. Also, (3.2) is independent of m. That is saying that when taking $m \to +\infty$, $L(\vec{\alpha})$ stays the same. Therefore, maximizing the true likelihood with respect to β is equivalent to maximizing (3.2) with respect to β. □