BAYESIAN ANALYSIS OF SEMIPARAMETRIC PROPORTIONAL HAZARDS MODELS

A.E. Gelfand
B.K. Mallick

TECHNICAL REPORT No. 479
MARCH 21, 1994

Prepared Under Contract
N00014-92-J-1264 (NR-042-267)
FOR THE OFFICE OF NAVAL RESEARCH

Professor David Siegmund, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305-4065

Bayesian analysis of semiparametric proportional hazards models

Alan E. Gelfand and Bani K. Mallick

Summary

We consider the usual proportional hazards model in the case where the baseline hazard, the covariate link and the covariate coefficients are all unknown. Both the baseline hazard and the covariate link are monotone functions and are characterized nonparametrically using a dense class arising as a mixture of Beta distribution functions. We take a Bayesian approach for fitting such a model. Since interest focuses more upon the likelihood, we consider vague prior specifications including Jeffreys's prior. Computations are carried out using sampling-based methods. Model criticism is also discussed. Finally, a data set studying survival of a sample of lung cancer patients is analyzed.

Key words: Bayesian model analysis; Gibbs sampler; Mixture-of-Betas model; model criticism; survival analysis.

1 Introduction

Probablistic models and statistical analysis for technological and medical survival data have garnered much attention over the past twenty years. Adopting the medical setting data is usually obtained for a sample of individuals. For the ith subject, data consists of either an observed survival time t_i or a (right) censorship time v_i (in which case $t_i > v_i$) and a set of explanatory variables or risk factors denoted by the p×1 vector $\mathbf{x}_i$.

Survival distributions are usually characterized by their hazard function which is the conditional density function at time t given survival up to time t. In customary modeling, the hazard for an individual is assumed to be a function of the covariates as well as t. To accommodate censored survival times we also require the survivor function which is the upper tail probability function of the survival distribution. When survival times are assumed to be continuous measurements, a widely used class of hazard functions takes the so-called proportional hazards form,

$$h(t) = h_0(t)g(\mathbf{x}) \tag{1}$$

with h_0 and g nonnegative.

The integrated hazard associated with $h(t)$ is $H(t) = H_0(t)g(\mathbf{x})$ where $H_0(t) = \int_0^t h_0(u)du$. The survivor function is $S(t;x) = exp(-H_0(t)g(\mathbf{x}))$ and the density for t takes the form

$$f(t;\mathbf{x}) = h_0(t)g(\mathbf{x})e^{-H_0(t)g(\mathbf{x})} \tag{2}$$

Implicit in (1) is the fact that the random variable $U(t;\mathbf{x}) = H_0(t)g(\mathbf{x})$ has an Exp(1) distribution. We shall make use of this later. The function h_0 is called the baseline hazard associated with a baseline survival distribution, i.e., when g=0. If for example, this baseline distribution is a Weibull, the associated h_0 has a parametric form $h_0(t) = \rho t^{\rho-1}$. The integrated baseline hazard, $H_0(t)$ as defined above is clearly a *nondecreasing* function of t.

Dating to Cox (1972), the function g is customarily taken to have the parametric form $exp(\boldsymbol{\beta}^T\mathbf{x})$ where $\boldsymbol{\beta}$ is an unknown $p \times 1$ vector of coefficients. Drawing upon jargon for generalized linear models, g can be more generally thought of as a covariate link function.

In fact, letting $\eta = \beta^T \mathbf{x}$, because $EU(t;\ \mathbf{x})=1$, $EH_0(t) = 1/g(\eta)$. Since $EH_0(t)$ increases as $E(t)$ increases and $E(t)$ decreases in $g(\eta)$ we have that g is a *non decreasing* function of η. However since $\beta^T \mathbf{x}$ can include polynomial forms for any covariate, g need not be a monotonic function of any covariate.

We shall assume that H_0 is an unknown function from $R^+ \rightarrow R^+$ and that g is an unknown function from $R^1 \rightarrow R^+$ which depends upon $\mathbf{x}$ through $\eta = \beta^T \mathbf{x}$ with β an unknown parameter vector. Nonparametric estimation of h_0, hence H_0, usually proceeds from a piecewise constant form for h_0 as in Breslow (1974). See Cox and Oakes (1984) for a more general discussion. If g is specified and $\hat{\beta}$ is obtained by partial likelihood maximization (Cox, 1975) then $\hat{h_0}$ and $\hat{H_0}$ are determined. If the full likelihood is used maximization is carried out iteratively, first for h_0 given β, then for β given h_0, etc., until convergence.

Conventional (wisdom offered for instance in Cox and Oakes, 1984 or McCullagh and Nelder, 1989) asserts that if h_0 is estimated nonparametrically, the overall fit is insensitive to the choice of $g(\eta)$, supporting the usuage of the mathematically convenient form e^{η}. However nonparametric estimation of g has recently received attention. For instance, in O'Sullivan (1988) and in Hastie and Tibshirani (1990), it is assumed that $g(\mathbf{x}) = e^{\theta(\mathbf{x})}$ where $\theta(\mathbf{x})$ is an arbitrary additive function in the components of $\mathbf{x}$, i.e., $\theta(\mathbf{x}) = \sum_{l=1}^{p} \theta_l(x_l)$. Partial likelihoods are used so that g is estimated independently of H_0. We note that the form $e^{\theta(\mathbf{x})}$ does not include our assumed form $g(\beta^T \mathbf{x})$. Moreover, an estimate of $g(\eta)$ is easily compared with e^{η}, which we may think of as a "baseline" covariate link. Staniswalis (1989) considers joint estimation of unknown H_0 and g with only a single dichotomous covariate. Indeed if $\mathbf{x}$ is a single continuous covariate, i.e., $\eta = \beta_0 + \beta_1 x$, then H_0, g, β_0 and β_1 need not be identifiable in the likelihood. Staniswalis's approach uses kernel estimators resulting in the maximization of a weighted likelihood. Her fitting algorithm is also iterative, maximizing over H_0 given g, then over g given a new H_0, etc.

We model H_0 and g using a dense class of monotone functions from R^+ to R^+ and from R^1 to R^+ respectively. The classes arise from mixtures of Beta distribution functions and have not been previously considered in this context. We adopt a Bayesian framework to

fit such a model thus assuming β, H_0 and g are random. Bayesian fitting yields an entire posterior distribution for a model unknown rather than merely a point estimate and perhaps a precision estimate. Inference is exact rather than asymptotic. If useful prior information is available say about coefficient parameters we would be happy to use it. However, interest usually focuses more upon the likelihood whence we would tend to use noninformative priors (see section 3 and the example of section 5.). With large data sets such automatic Bayesian inference will be close to that arising under classical maximum likelihood analysis. For smaller data sets the Bayesian approach would be expected to provide more believable estimates of variability than under likelihood analysis. Recent advances in Bayesian computation through the use of sampling based methods (Gelfand and Smith, 1990; Smith & Gelfand, 1993) enable reasonably straightforward fitting of such models. Bayesian fitting of fully parametric proportional hazards models using sampling based methods is discussed in Dellaportas and Smith (1992).

Our approach models the strictly monotone functions H_0 and g, each transformed by a suitable monotone transformation to have range in [0,1], as unknown cumulative distribution functions. In the process the resulting domain also becomes [0,1]. We thus characterize each function as a mixture of Beta distribution functions. Here we appeal to a well known result as in, e.g., Diaconis and Ylvisaker (1985), which says that any continuous density on [0,1] can be arbitrarily well approximated by a discrete mixture of Beta densities. Unlike distributions arising under a Dirichlet process, which could also be used here, we have a continuous, dense class of distributions admitting an explicit form. In practice we have treated the number of mixands r, as fixed comparing various choices. Alternatively, a discrete prior could be attempted. As such we have taken an infinite dimensional problem and converted it to a finite dimensional one. Though arbitrarily high dimensional models can be employed, happily, in practice they will not usually be needed. In experimenting with a range of r's, for a number of examples, we have discovered, perhaps not surprisingly, that robustness occurs with quite small r. In introducing randomness to this finite mixture form we could either assume the mixture weights to be random or the parameters of the Beta densities to be random. Mathematically it is much simpler to work with the former. Hence for a given r we

choose a fixed set of Beta densities and then assume, a priori, that the mixing weights arise as a random draw on the r-dimensional simplex. Earlier work with this approach (Mallick and Gelfand, 1993) looked at the problem of fitting a generalized linear model when the link function is assumed unknown.

The format of the paper is thus as follows. In section 2 we formalize the likelihood, in particular the modeling of the unknown H_0 and g. In section 3 we discuss the prior specification needed to complete the Bayesian model. Section 4 describes briefly the sampling based fitting of this model. Model criticism and comparison are the subject of section 5. Finally, in section 6 we analyze a set of survival data, with censoring, on 40 advanced lung cancer patients which is discussed in Lawless (1982).

2 Likelihood specification

Suppose a total of n subjects. For the ith individual let $\gamma_i = 1$ if $t_i > v_i$, =0 otherwise and let $y_i = min(t_i, v_i)$. Then the joint density of the sample, following from (2), is

$$\prod_{i=1}^{n} [h_0(y_i)g(\beta^T \mathbf{x_i})]^{1-\gamma_i} exp(-H_0(y_i)g(\beta^T \mathbf{x_i})) \tag{3}$$

With H_0, g and β unknown we may view (3) as a likelihood function $L(\beta, H_0, g)$. Indeed L has an infinite dimensional argument; without further assumptions it need not be identifiable. For example, if $\beta^T \mathbf{x} = \beta_0 + \beta_1 x$ then H_0, g, β_0, and β_1 cannot be identified.

As noted in section 1, our inference approach is Bayesian requiring the specification of a prior $f(\beta, H_0, g)$ whence $L(\beta, H_0, g) \cdot f(\beta, H_0, g)$ is the Bayesian model. In section 3 we discuss forms for the prior f. The identifiability question from a Bayesian point of view becomes whether or not the data can inform about all of the unknown parameters in the model. If yes, provided a proper posterior results, there is no identifiability problem. If no and if f is improper then the posterior necessarily is as well and we have an ill-defined Bayesian model. If no and if f is proper then the prior drives the posterior.

How shall we model H_0 and g? From the previous section we assume H_0 is a strictly

increasing function from R^+ onto R^+ and that g is a strictly increasing function from R^1 onto R^+. Let $J_0 = a_0 H_0/(a_0 H_0 + b_0)$ where $a_0, b_0 > 0$ are specified constants (choice of a_0, b_0 will be taken up below). Then J_0 is a differentiable c.d.f. Modeling J_0 is equivalent to modeling an unknown distribution function. Similarly, let $J_g = a_1/(a_1 g + b_1)$ where $a_1, b_1 > 0$ (choice of a_1, b_1 will be clarified below). Then again J_g is a differentiable c.d.f.

A rich class of models for J_0 and for J_g may be created as follows. Associate with H_0 a specific baseline hazard function $\tilde{H}_0$. For instance $\tilde{H}_0(t)$ might be t^ρ for a given ρ. (Random ρ could be handled using a hyperprior but we have not investigated this). Similarly, associate with g a specific baseline covariate link $\tilde{g}$. For instance $\tilde{g}(\eta)$ might be e^η. Diaconis and Ylvisaker (1985) argue that discrete mixtures of Beta densities provide a continuous dense class of models for densities on [0,1]. A member of this class has a density of the form

$$f(u) = \sum_{l=1}^{r} w_l Be(u|c_l, d_l) \tag{4}$$

where r denotes the number of mixands, $w_l \geq 0$, $\sum w_l = 1$ and $Be(u|c_l, d_l)$ denotes the Beta density in standard form with parameters, c_l and d_l. If $IB(u; c_l, d_l)$ denotes the incomplete Beta function associated with $Be(u|c_l, d_l)$ then let

$$J_0(t; \mathbf{w}_1) = \sum_{l=1}^{r_1} w_{1l} IB(\tilde{J}_0(t); c_{1l}, d_{1l}) \qquad \text{and} \qquad J_g(\eta; \mathbf{w}_2) = \sum_{l=1}^{r_2} w_{2l} IB(\tilde{J}_g(\eta); c_{2l}, d_{2l}) \tag{5}$$

where $\tilde{J}_0(t) = a_0 \tilde{H}_0(t)/(a_0 \tilde{H}_0(t) + b_0)$ and $\tilde{J}_g(\eta) = a_1 \tilde{g}(\eta)/(a_1 \tilde{g}(\eta) + b_1)$. Clearly J_0 and J_g are c.d.f.'s.

It will be of interest in subsequent sections to calculate $h_0(t) = H_0'(t)$ and $g'(\eta)$. Since $H_0 = b_0 J_0/a_0(1 - J_0)$, $H_0' = (b_0\, \partial J_0/\partial t)/a_0(1 - J_0)^2$. Similarly $g = b_1(1 - J_g)/a_1 J_g$ so $g' = (-b_1\, \partial J_g/\partial \eta)/a_1 J_g^2$. From (5), $\partial J_0(t; \mathbf{w}_1)/\partial t = \tilde{J}_0' \sum_{l=1}^{r_1} w_{1l} Be(\tilde{J}_0(t); c_{1l}, d_{1l})$, $\partial J_g(\eta; \mathbf{w}_2)/\partial \eta = \tilde{J}_g' \sum_{l=1}^{r_2} w_{2l} Be(\tilde{J}_g(\eta); c_{2l}, d_{2l})$ with $J_0' = a_0 b_0 \tilde{H}_0'/(a_0 \tilde{H}_0 + b_0)^2$ and $J_g' = -a_1 b_1 \tilde{g}_0'/(a_1 \tilde{g} + b_1)^2$.

We note that models incorporating mixtures other than Beta could be used, e.g., Gammas on R^+, uniforms on R^1. We work with Beta densities since the restriction to the bounded interval [0,1] enables, for a given r, convenient choice of the set of c_l, d_l so that mixing with a weight vector w=($w_1, \cdots w_r$) belonging to the r-dimensional simplex yields a rich family of densities.

In our experience, inference using mixtures with small r is virtually indistinguishable from those with much larger r. In fact, allowing r≥n does not insure perfect fit since H_0 and g are restricted to be monotone. Hence we specify r making (3) a finite dimensional likelihood. Given r, it is natural and certainly mathematically easier to assume that the component Beta densities are specified but that the weights are unknown. In particular we choose the set of c_l, d_l to provide a collection of Beta densities which blanket [0,1]. Within this objective, our experience shows that inference is again robust to the particular choice. In our example we set $r_1 = r_2 = r, c_{1l} = c_{2l} = c_l, d_{1l} = d_{2l} = d_l$ with c_l=λl, $d_l = \lambda$(r+1-l). (See section 3 for further discussion of this point.) In any event, specification of H_0 and g is equivalent to specification of w_1 and w_2 and we can denote the likelihood by $L(\beta, \mathrm{w}_1, \mathrm{w}_2)$.

Lastly we return to the choice of a_0, b_0 and a_1, b_1. This is not a modeling issue but a computational one. For example, if $|\eta|$ is very large, $\bar{g}(\eta) = e^{\eta}$ will produce over or underflow. Since, in (5), only $\tilde{J}_g(\eta)$ is needed, scaling and centering using appropriate a_1 and b_1 can alleviate this problem. In practice a_1 and b_1 could be obtained by looking at the range of $\tilde{\beta}^T \mathrm{x}$ over the sample with $\tilde{\beta}$ say the partial likelihood maximizer. For $\tilde{J}_0(t)$, often $a_0 = b_0$=1 will work. Again, there is no notion of "best" values. Many choices of a_i and b_i will work equally well and, to keep notation simpler, we suppress the a_i and b_i in the sequel.

3 Prior specification

Given the likelihood $\mathrm{L}(\beta, \mathrm{w}_1, \mathrm{w}_2)$ we next address the specification of the prior $f(\beta, \mathrm{w}_1, \mathrm{w}_2)$. Since primary interest is in the likelihood and since only occasionally will there be useful prior information we think in terms of vague priors. In this regard one advantage does accrue to the Bayesian compared with the likelihoodist. Maximum likelihood estimators under $\mathrm{L}(\beta, \mathrm{w}_1, \mathrm{w}_2)$ need not be finite, i.e., need not occur within the interior of the parameter space (see Wedderburn, 1976). However in the Bayesian context the prior distribution typically overcomes this problem yielding a well behaved posterior density.

We assume that $f(\beta, \mathrm{w}_1, \mathrm{w}_2)$= $f(\beta|\mathrm{w}_1, \mathrm{w}_2).f(\mathrm{w}_1).f(\mathrm{w}_2)$ where $f(\mathrm{w}_i)$ is a distribution

on the r_i dimensional simplex. Since $\mathbf{w}_1$ determines $H_0(t)$ we might attempt to choose $f(\mathbf{w}_1)$ such that in some senses $H_0(t)$ is *centered* around $\tilde{H}_0(t)$, i.e., $J_0(t;\mathbf{w}_1)$ is *centered* around $\tilde{J}_o(t)$. From (4) this corresponds to centering the random density $f(u)$ around a uniform density on [0,1]. Centering using the mean results in the equation

$$\sum_{l=1}^{r_1} E(w_{1l} IB(u; c_{1l}, d_{1l}) = u \tag{6}$$

If $\mathbf{w}_1 \sim Dir(\gamma \mathbf{1})$ then (6) requires that the average of the $IB(u; c_{1l}, d_{1l})$ equal u. If r_1 is even and c_{1l} and d_{1l} are chosen as discussed in section 2 then straightforward calculation (which we omit) shows that, to a first order approximation, (6) holds. Similar remarks apply to $f(\mathbf{w}_2)$ with regard to *centering* $g(\eta)$ around $\tilde{g}(\eta)$.

Returning to $f(\boldsymbol{\beta}|\mathbf{w}_1, \mathbf{w}_2)$ a multivariate normal prior is often proposed. In the limit, as the precision matrix tends to 0, a flat prior results. It is interesting to investigate the form of Jeffreys's prior here. Recall that this prior is the square root of the determinant of the Fisher information matrix associated with $L(\boldsymbol{\beta}, \mathbf{w}_1, \mathbf{w}_2)$. Given $\mathbf{w}_1$ and $\mathbf{w}_2$, standard calculation shows that Jeffreys's prior is proportional to $|\mathbf{X}^T M \mathbf{X}|^{1/2}$ where $\mathbf{X}$ is the $n \times p$ matrix whose rows are the $\mathbf{x}_i^T$'s and M is an $n \times n$ diagonal matrix such that

$$M_{ii} = -(1 - S(v_i; \mathbf{x}_i))(g'(\boldsymbol{\beta}^T \mathbf{x}_i)/g(\boldsymbol{\beta}^T \mathbf{x}_i))^2 \tag{7}$$

Working with (3), this calculation is clarified by noting that $E(1 - \gamma_i - H_0(y_i) g(\boldsymbol{\beta}^T \mathbf{x}_i)) = 0$. From (7) we see that Jeffreys's prior depends upon both $\mathbf{w}_1$, $\mathbf{w}_2$. Recall that g' was calculated in section 2; g'' follows similarly. In the case of of no censoring, $M_{ii} = -(g'(\boldsymbol{\beta}^T \mathbf{x}_i)/g(\boldsymbol{\beta}^T \mathbf{x}_i))^2$ which depends only upon $\mathbf{w}_2$. Finally, the question of whether a proper posterior results under the above specifications can be examined using ideas in Ibrahim and Laud (1991). No details are given here.

4 Implementing the Bayesian model.

Given the Bayesian model $L(\boldsymbol{\beta}, \mathbf{w}_1, \mathbf{w}_2) \cdot f(\boldsymbol{\beta}|\mathbf{w}_1, \mathbf{w}_2) \cdot f(\mathbf{w}_1) \cdot f(\mathbf{w}_2)$, all inference proceeds from the posterior distribution of $(\boldsymbol{\beta}, \mathbf{w}_1, \mathbf{w}_2)$ which is proportional to this product. Analytic investigation of this $p + (r_1 - 1) + (r_2 - 1)$ dimensional nonnormalised joint distribution

is infeasible. It would be difficult enough to standardize this form much less to calculate expectations, marginal distributions, etc. Thus we adopt a sampling based approach using a Markov chain Monte Carlo algorithm to obtain random draws essentially from this posterior distribution. In particular, a version of the Gibbs sampler (Gelfand and Smith, 1990) is natural since the complete conditional distributions for $\beta, \mathbf{w}_1$ and $\mathbf{w}_2$ are also proportional to the nonnormalized posterior. Because, for a given β, calculation of $L(\beta, \mathbf{w}_2, \mathbf{w}_1)$ requires $n(r_1 + r_2)$ incomplete Beta function evaluations, making draws directly from these distributions is inconvenient. Instead we utilise a Metropolis-within-Gibbs algorithm (Müller, 1993) with a multivariate normal proposal density for β, and Dirichlet proposal densities for $\mathbf{w}_1$ and $\mathbf{w}_2$. Under a logit transformation of $\mathbf{w}_i$ to $r_i - 1$ dimensional space these last proposals could be multivariate normal. We run short Metropolis sub-chains, typically 20 to 50 iterations, for each update within each iteration of the Gibbs sampler. Such sub-chains are attractive in that, to proceed from the current step to the next requires only one new evaluation of the likelihood. Starting points for replications of the Gibbs sampler are taken in the vicinity of the maximum likelihood estimates for β under $\tilde{H}_0$ and $\tilde{g}_0$ with random uniform draws for the $\mathbf{w}_i$. In the example of section 5 we used 1000 parallel chains initially, adaptively improving the proposal densities, following Müller's suggestions, for typically 25 iterations after which we ran ten parallel Gibbs chains each for approximately 5000 iterations using various "convergence" diagnostics before stopping.

Let us denote the retained output of the Gibbs sampler, which is approximately a sample from the posterior, by $(\beta_j^*, \mathbf{w}_{1j}^*, \mathbf{w}_{2j}^*)$ $j = 1, 2, \ldots, m$ where typically m is 1000 to 2000. This sample enables us to carry out any desired posterior inference. Such inference would likely examine the marginal posterior distributions of the coefficients. The posteriors of H_0 and h_0 can be obtained at any t using the posterior of $\mathbf{w}_1$. The posterior of g can be obtained at any η using the posterior of $\mathbf{w}_2$.

5 Model criticism and comparison

Here we consider informal techniques for checking model assumptions and for comparing models. In the non Bayesian framework, with censored data, assessment of the hazard function is usually done using a hazard plot. More specifically, a parametric specification for H_0, equivalently h_0, can be checked using an empirical estimate of H_0 or h_0. Plots of $\hat{H}_0$, $\log\hat{H}_0$, $\hat{h}_0$ or $\log\hat{h}_0$ vs t or $\log t$ can be compared with theoretical plots under the parametric specification. An alternative is a *residual* analysis based upon the earlier remark that given β, H_0 and g, $U_i = U(t_i; \mathbf{x}_i) = H_0(t_i)g(\beta^T \mathbf{x}_i) \sim \text{Exp}(1)$. Taking H_0 and g as known and inserting a sample estimate for β, the resulting $\hat{U}_i$ are compared with the Exp(1) distribution using a theoretical or empirical QQ plot (see, e.g., Lawless, 1982; or Cox and Oakes, 1984). Unfortunately the required estimation sacrifices both the independence and exact distribution for the U_i.

How should such checking be handled within the Bayesian framework? Generally, model assessment proceeds from predictive distributions (Box, 1980). At the individual level this proposes, for the observed survival times, comparison of $f(t_i|data)$ with $t_{i,obs}$; for the censored survival times evaluation of $S(t_i|data)$ at v_i. It may be preferable not to include the observed information on the ith individual in obtaining the predictive distribution for t_i. That is, we would condition on $data(i)$, all of the data excluding the information on the ith individual. (Such deletion is usually referred to as crossvalidation.) Regardless, Gelfand and Dey (1993) describe how to use the output of the Gibbs sample to obtain Monte Carlo estimates of either of these predictive densities and predictive survivor functions.

For instance $\hat{f}(t_i|data) = m^{-1}\sum_{j=1}^{m} f_j^*(t_i; x_i)$ where $f_j^*(t_i; x_i)$ denotes the density (2) given $(\beta_j^*, \mathbf{w}_{1j}^*, \mathbf{w}_{2j}^*)$. Similarly $\hat{S}(t_i|data) = m^{-1}\sum_{j=1}^{m} S_j^*(t_i; x_i)$ where $S_j^*(t_i; x_i)$ denotes the survivor function associated with (2) given $(\beta_j^*, \mathbf{w}_{1j}^*, \mathbf{w}_{2j}^*)$. Appropriate resampling of the $(\beta_j^*, \mathbf{w}_{1j}^*, \mathbf{w}_{2j}^*)$ enables conditioning on $data(i)$ (Smith and Gelfand, 1992). In particular for an observed t_i we obtain the conditional predictive ordinate (CPO) as $\hat{f}(t_{i,obs}|data(i))$. For an unobserved t_i we can obtain $\hat{S}(v_i|data(i))$. A large CPO indicates agreement between the observation and the model (see Pettit and Young, 1990). Hence models can be compared

using a plot vs. i of, e.g., CPO's under the various models, CPO ratios, or log CPO ratios. Similarly, for the censored observations, a large value of $\hat{S}(v_i|data(i))$ indicates agreement between the observed censoring and the model.

6 An illustrative example

The data in table 1 describes survival data in days for 40 advanced lung cancer patients and is taken from Lawless (1982). The three explanatory variables are : x_1, performance status at diagnosis (a measure on a scale from 0 to 100 of the patient's general medical condition), x_2, age of the patient in years, and x_3, months from diagnosis to entry into the study. Note that three of the 40 survival times, indicated by * are censored. An exponential model, i.e., $h_0(t)$=1, $g(\eta) = e^{-\eta}$, is used by Lawless. For this four parameter model (intercept and three coefficients) maximum likelihood estimates and associated standard errors can be obtained from standard statistical packages and are presented in table 2. Accordingly, only x_1 is very important. For the 37 uncensored survival times, figure 1 provides a theoretical Q-Q plot for the $\hat{U}_i = t_{i,obs} exp(-\hat{\beta}^T \mathbf{x}_i)$. The exponential model appears adequate but would we prefer a semiparametric choice?

We investigated four models from a Bayesian perspective. Model 1 has a likelihood arising under an exponential hazard and an exponential covariate link and is denoted by EE. Model 2 has a nonparametric hazard specification with an exponential covariate link and is denoted by NE. The integrated hazard is developed as in section 2 "centered" about the exponential hazard. Model 3 has an exponential hazard but a nonparametric covariate link "centered" about the expoponential function and is denoted by EN. Finally model 4 incorporates nonparametric hazard and covariate link each exponentially centered and denoted by NN. When nonparametric forms are used we took $\lambda = 1$ and $r_i = 3$. We investigated larger r_i's but witnessed inconsequential improvement in fit primarily because, regardless of r_i, the function is constrained to be monotonic. This is an obvious advantage to our modeling approach. We can fit H_0 and g using a large number of mixands but, in practise, a much more parsimonious choice can usually be taken. We used a flat prior on

β and, if $\mathbf{w}_i$ appears we chose a uniform prior on the simplex in three dimensions. Hence model 1 has 4 parameters, models 2 and 3 have 6 and model 4 has 8.

With regard to model choice, in figure 2 we present a CPO plot for each of the four models. The three semiparametric choices are better than EE with the largest model, NN, the best. For each of the three censored survival times under each model, Table 3 gives $P(t_i > v_i|data(i))$ along with the product over these three times. Models NE and NN are best at explaining the censoring.

Looking at model choice in a different way suppose we compare the estimated nonparametric functions with their centering exponentials. As the Bayes estimate of the nonparametric function we take the posterior mean. In figure 3a we compare the integrated hazards under NE and under NN with $H_0(t) = t$. In figure 3b we compare the hazards under NE and under NN with $h_0(t) = 1$. The domain for t agrees in both figures and includes virtually all the observed t_i. We note that $\hat{H}_0$ and $\hat{h}_0$ are almost indistinguishable under NE and NN but differ dramatically from those for the exponential model. Figure 3 shows that for this data the aforementioned claim of Cox and Oakes (1984) and of McCullagh and Nelder (1989) holds: if H_0 is estimated, the fit is insensitive to the covariate link. In figure 4 we compare the covariate links under EN and NN with $g(\eta) = e^{\eta}$. It is convenient to use the ratio $\hat{g}/e^{\eta}$ plotted over a range which includes all of the $\hat{\beta}^T x_i$. The $\hat{g}$'s under EN and under NN are in good agreement for $\eta > 0$ with growing disagreement as η decreases.

All of the above discussion supports NE or NN. Looking at NN in more detail we find the posterior mean and standard deviations for the β_i under NN in table 2. In terms of inference for β alone there is little qualitatative difference between EE and NN. Finally, in table 4 we provide the posterior correlations amongst the 8 parameters under NN. As expected, there is high negative correlation between w_{11} and w_{12} and between w_{21} and w_{22}. Otherwise, correlations are weak and $\mathbf{w}_1$ (i.e. H_0) is essentially uncorrelated with $\mathbf{w}_2$ and β (i.e., g).

In conclusion, though the present data set does not, perhaps, warrant a more elaborate model than EE, it is attractive to have a broader model formulation and the associated fitting and choice tools to demonstrate this.

References

[1] Box, G.E.P (1980), Sampling and Bayes inference in scientific modeling and robustness. (with discussion). *Journal of the Royal Statistical Society, Series A,* **143**, 382-430.

[2] Breslow, N. (1974), Covariate analysis of censored survival data. *Biometrics* **30**, 89-99.

[3] Chaloner, K. and Brant, R. (1988), A Bayesian approach to outlier detection and residual analysis. *Biometrika* **75**, 651-659.

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

[5] Cox, D.R. (1975), Partial likelihood. *Biometrika* **62**, 269-276.

[6] Cox, D.R. and Oakes, D. (1984), *Analysis of Survival Data.* Chapman and Hall, London.

[7] Dellaportas, P. and Smith, A. F. M. (1992), Bayesian inference for generalised linear and proportional hazards models via Gibbs sampling. *Applied Statistics* 42, 443-460.

[8] Diaconis, P. and Ylvisaker, D. (1985), Quantifying prior opinion. In:*Bayesian Statistics* 2, eds. J.M.Bernardo et al 133-156. North Holland, Amsterdam.

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

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

[11] Gelfand, A.E. and Dey, D.K. (1993). Bayesian model choice: asymptotics and exact calculations. To appear in *Journal of the Royal Statistical Society, series B* (To appear).

[12] Hastie, T. and Tibshirani, R. (1990) Exploring the nature of covariate effects in the proportional hazard model. *Biometrics*, **46**, 1005-1016.

[13] Ibrahim. J, and Laud, P. (1991). On Bayesian analysis of generalized linear models using Jeffreys's prior. *Journal of the American Statistical Association*, **86**, 981-986.

[14] Kalbfleisch, J. D. (1978) Nonparametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society, Series B,* 40 214-221.

[15] Lawless, J. F. (1982) *Statistical models and methods for lifetime data.* John Wiley and Sons, New York.

[16] Lehmann, E. (1986) *Testing Statistical Hypothesis.* John Wiley and Sons, New York.

[17] Mallick, B. K. and Gelfand, A. E. (1993) Generalized linear models with unknown link functions. *Biometrika* (To appear).

[18] McCullagh. P. and Nelder. J. A. (1989), *Generalized linear models.* Chapman and Hall, London.

[19] Müller, P. (1991), A generic approach to posterior integration and Gibbs sampling. *Statistics and Computing* (to appear).

[20] O'Sullivan, F. (1988), Nonparametric estimation of relative risk using splines and cross-validation. *SIAM Journal on Scientific and Statistical Computing,* 9, 531-542.

[21] Pettit, L. I. and Young, K. D. S. (1990), Measuring the effect of observation on Bayes factors. *Biometrika,* 77, 455-466.

[22] Sinha, D. (1993) Semiparametric Bayesian analysis of multiple event time data. *J. Amer. Statist. Assoc.* (To appear)

[23] Smith, A. F. M. and Gelfand, A. E.(1992), Bayesian Statistics without tears : a sampling resampling perspective. *Journal of the American Statistical Association,* 46, 84-88.

[24] Staniswalis, J (1989) The kernel estimate of a regression function in likelihood-based models. *Journal of the American Statistical Association* 84, 276-283.

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

[26] Wedderburn, R. M. (1976), On the existence and uniqueness of the maximum likelihood estimates for certain generalised linear models. *Biometrika,* 63, 27-32.

Table 1: Lung Cancer Survival Data (Lawless, 1982)
* indicates censored observation

t	x_1	x_2	x_3	t	x_1	x_2	x_3
411	70	64	5	100	60	37	13
126	60	63	9	999	90	54	12
118	70	65	11	231*	50	52	8
92	40	69	10	991	70	50	7
8	40	63	58	1	20	65	21
25*	70	48	9	201	80	52	28
11	70	48	11	44	60	70	13
54	80	63	4	15	50	40	13
153	60	63	14	103*	70	36	22
16	30	53	4	2	40	44	36
56	80	43	12	20	30	54	9
21	40	55	2	51	30	59	87
287	60	66	25	18	40	69	5
10	40	67	23	90	60	50	22
8	20	61	19	84	80	62	4
12	50	63	4	164	70	68	15
177	50	66	16	19	30	39	4
12	40	68	12	43	60	49	11
200	80	41	12	340	80	64	10
250	70	53	8	231	70	67	18

Table 2: Summary of Likelihood Analysis for EE
Posterior Analyis for NN

	EE mle(s.e.)	NN post mean (s.d.)
β_0	4.742(.1612)	3.932(.1668)
β_1	0.060(.009)	0.035(.012)
β_2	0.013(.015)	0.001(.018)
β_3	.003(.010)	0.003(.009)

Table 3: Comparison Amongst Models of Survival Probabilities for Censored Observations

Model	(1) $P(t_1>25 \mid data(1))$	(2) $P(t_{23}>23 \mid data(23))$	(3) $P(t_{29}>103 \mid data(29))$	(1)(2)(3)
EE	0.8852	0.0253	0.5435	0.0121
NE	0.9091	0.1893	0.7019	0.1208
EN	0.8899	0.0904	0.6252	0.0503
NN	0.9133	0.1894	0.7116	0.1228

Table 4: Posterior Correlations under Model NN

	β_0	β_1	β_2	β_3	w_{11}	w_{12}	w_{21}
β_1	.194						
β_2	.026	.039					
β_3	.101	.267	−.042				
w_{11}	.007	−.017	.011	.018			
w_{12}	.004	.011	.015	.019	−.051		
w_{21}	.332	.245	−.102	.130	.019	.018	
w_{22}	−.141	.310	−.160	−.236	−.706	−.015	−.813

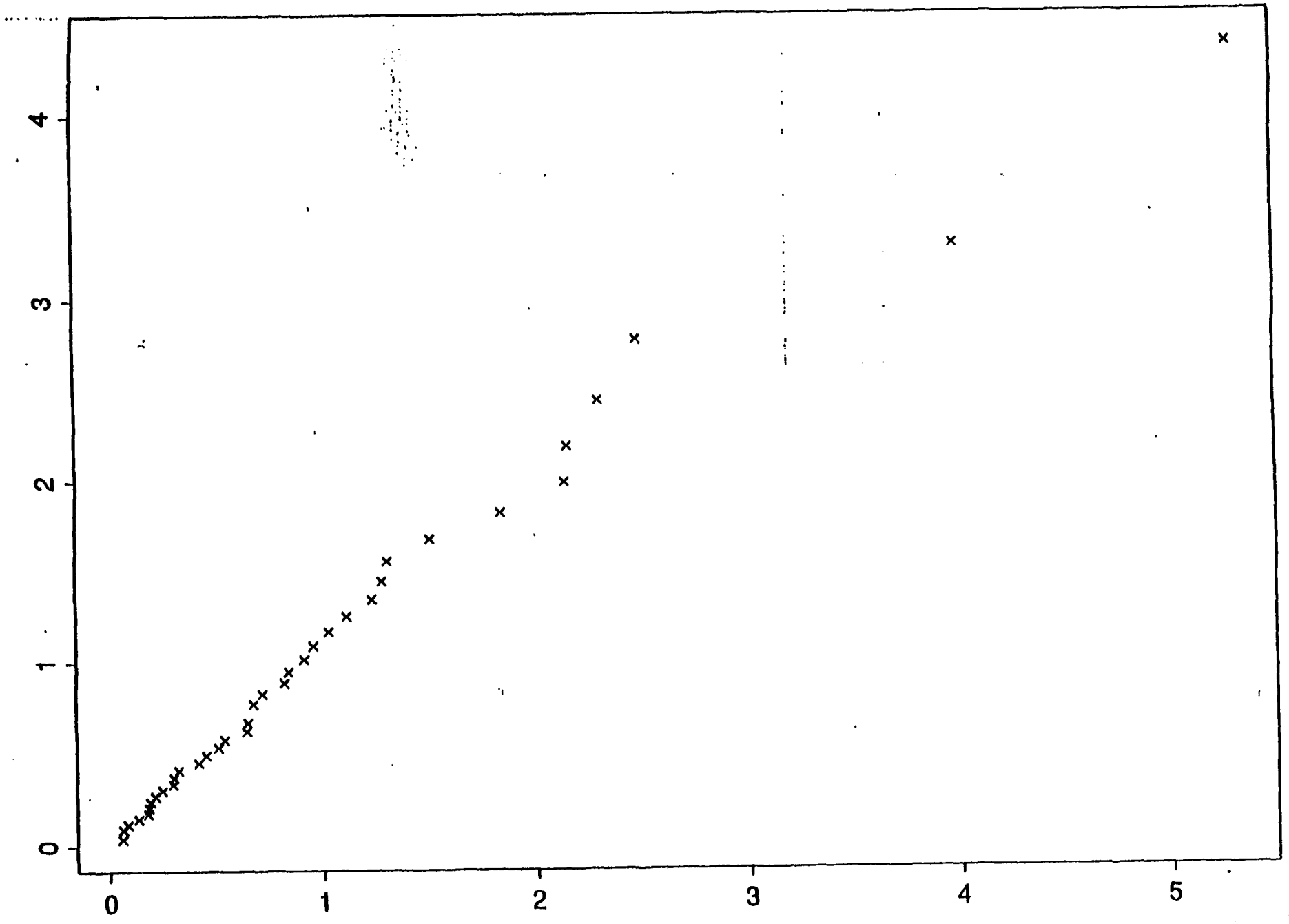

Figure 1: Theoretical Q-Q Plot for the $\hat{U}_i$ Under Exponential Model

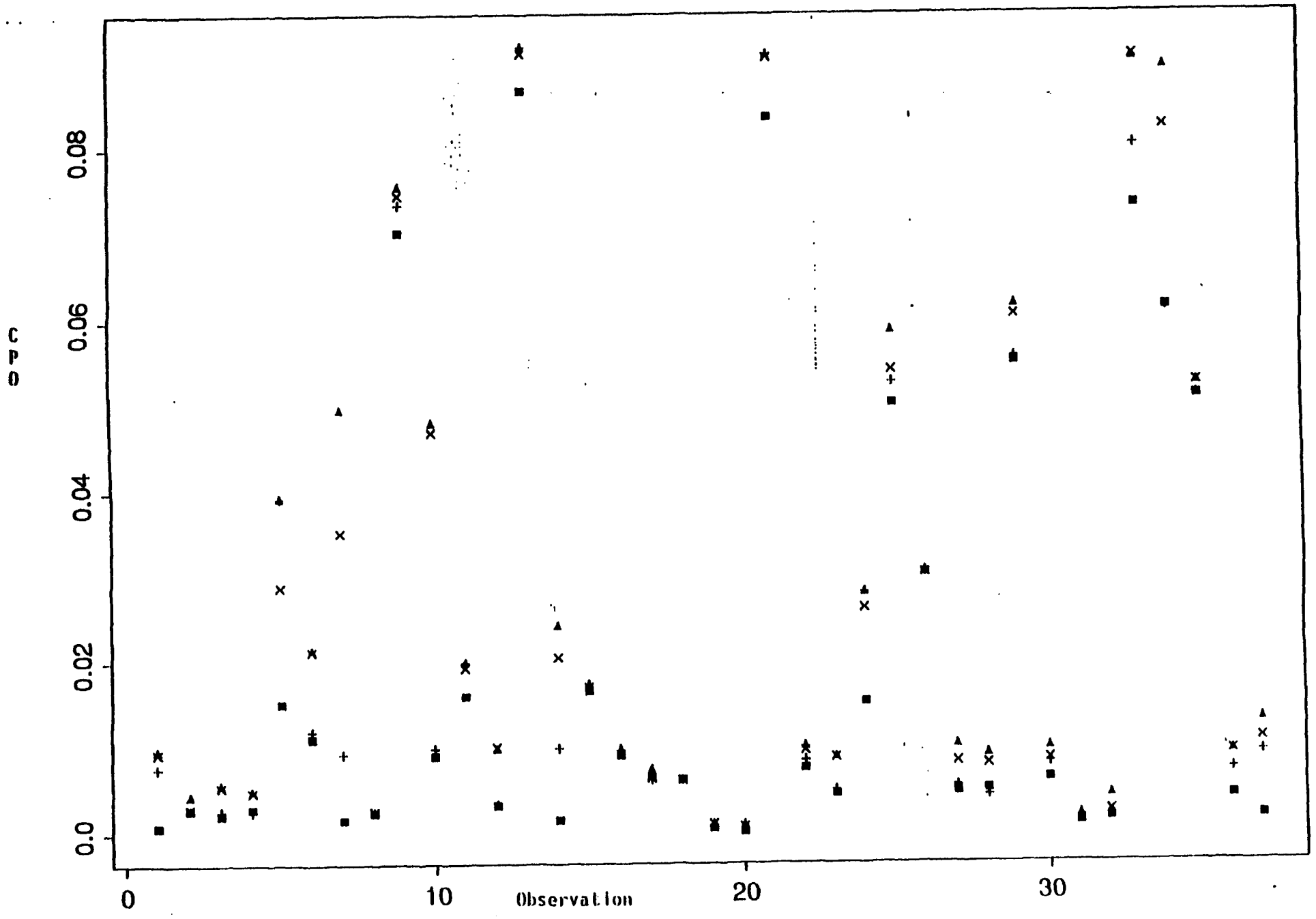

Figure 2: CPO Plot for Models EE([ ]), NE(x), EN(+) and NN(Δ)

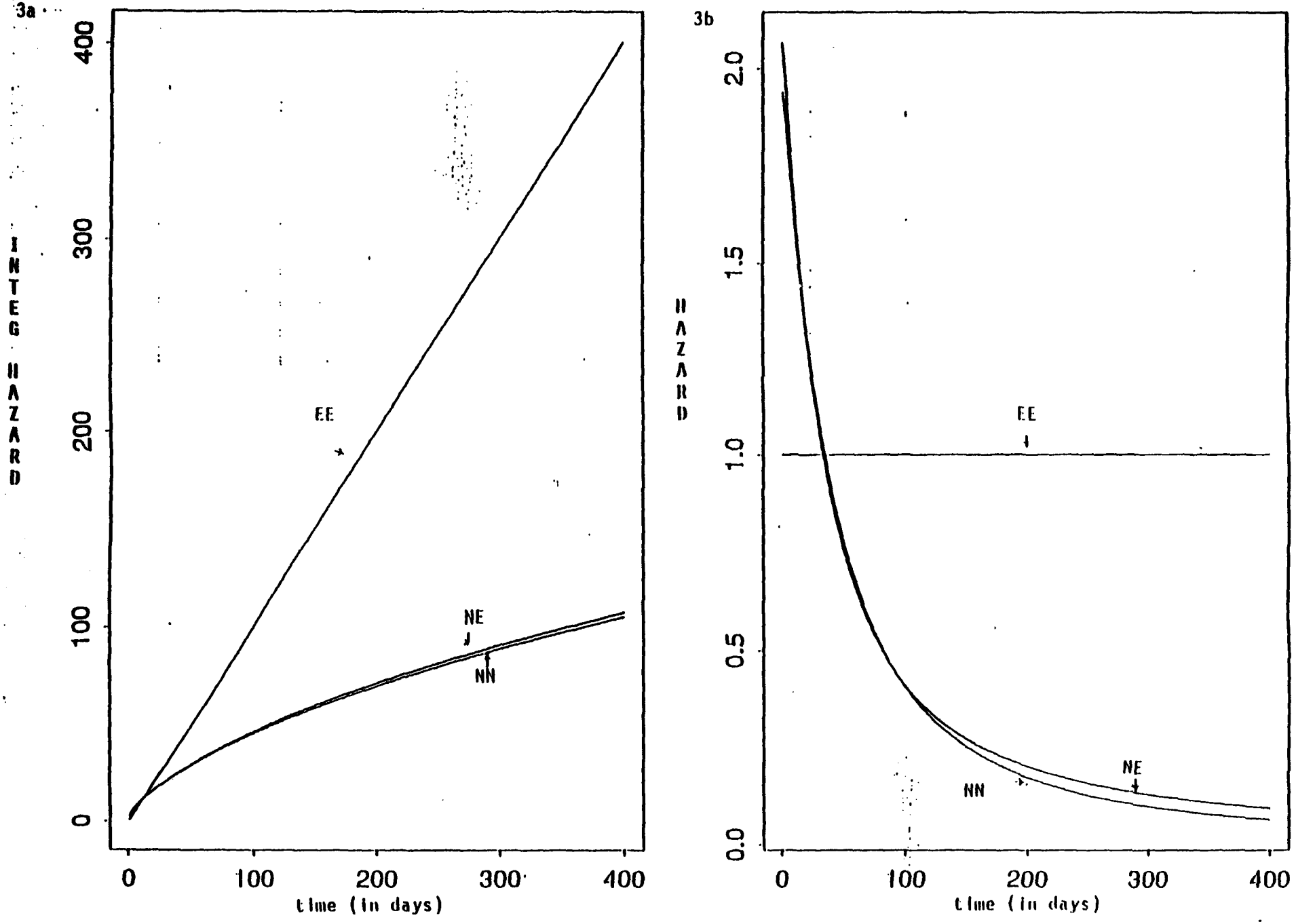

Figure 3: Comparison of Integrated Hazards (Figure 3a) and Hazards (Figure 3b) for models EE, NE and NN.

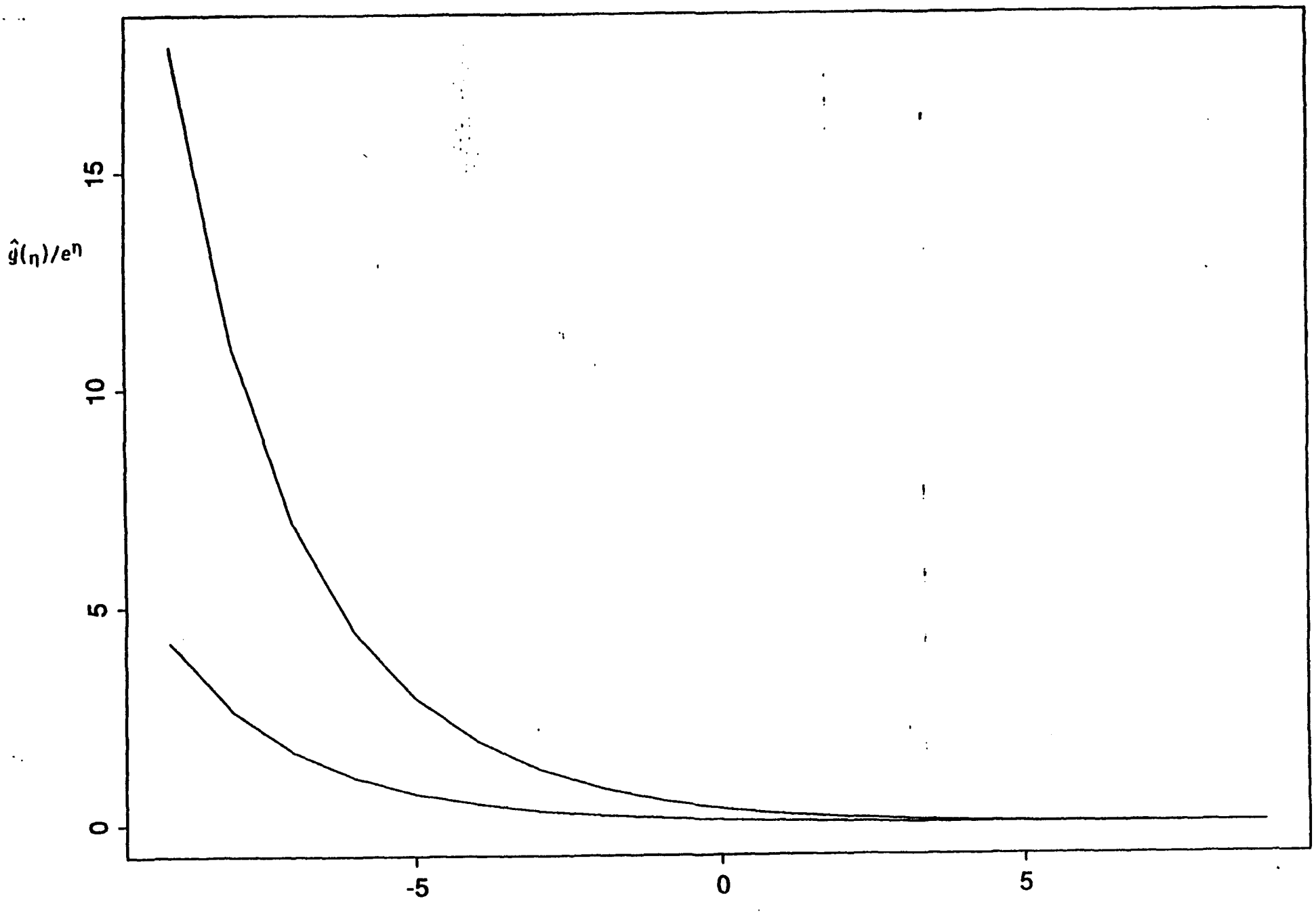

Figure 4: Comparison of Covariate Links for Models EE, EN and NN

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Bayesian Analysis of Semiparametric Proportional Hazards Models		5. TYPE OF REPORT & PERIOD COVERED Technical 6. PERFORMING ORG. REPORT NUMBER 479
7. AUTHOR(s) Alan E. Gelfand and Bani K. Mallick		8. CONTRACT OR GRANT NUMBER(s) N00014-92-J-1264
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Stanford University Stanford, CA 94305-4065		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR-042-267
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Statistics & Probability Program Code 1111		12. REPORT DATE 21 March 1994 13. NUMBER OF PAGES 20
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Bayesian model analysis; Gibbs sampler; Mixture-of-Betas model; model criticism; survival analysis

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

See reverse side

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. ABSTRACT

We consider the usual proportional hazards model in the case where the baseline hazard, the covariate link and the covariate coefficients are all unknown. Both the baseline hazard and the covariate link are monotone functions and are characterized nonparametrically using a dense class arising as a mixture of Beta distribution functions. We take a Bayesian approach for fitting such a model. Since interest focuses more upon the likelihood, we consider vague prior specifications including Jeffreys's prior. Computations are carried out using sampling-based methods. Model criticism is also discussed. Finally, a data set studying survival of a sample of lung cancer patients is analyzed.