Biometrika (2012), **99**, 1, *pp.* 1–11

Printed in Great Britain

Advance Access publication on 31 July 2012

Gaussian process regression for survival analysis with interval censored data

BY J. E. BARRETT AND A. C. C. COOLEN

Institute of Mathematical and Molecular Biomedicine, Hodgkin Building, King's College London, London SE1 1UL, U. K.

james.j.barrett@kcl.ac.uk ton.coolen@kcl.ac.uk

SUMMARY

We develop a new method for analysing survival data using Gaussian process regression to directly infer the relationship between event times (outputs) and covariates (inputs). We compare this to a second Gaussian process model that assumes a Cox-type hazard rate. Both of these approaches are applied to interval censored data, although they can easily be extended to accommodate any combination of left, right or interval censored or truncated observations. We define a general non-linear transformation model from which several existing models, including both our Gaussian process models, can be derived as special cases. Using either method hazard rates and survival curves can be extracted and we can perform variable selection. For censored individuals we can estimate what the event time would have been in the absence of censoring and noise. Results from simulated data illustrate that both models can infer non-monotonic relationships between the covariates and event times in the presence of right and interval censoring.

Some key words: Cox proportional hazards; Gaussian process; Interval Censoring; Survival analysis; Transformation model.

1. INTRODUCTION

We approach the analysis of time to event data as a regression problem. We want to model the event times as a stochastic function of the covariates. An elegant and powerful non-parametric method of doing this is Gaussian process regression (Rasmussen & Williams, 2006). This allows for flexible inference of a wide range of non-linear functions by specifying different kernel functions. The motivation behind this paper is to apply Gaussian process regression to censored observations. We will follow two alternative routes, one which focuses on the event time distribution and one which focuses on the hazard rate.

In the first route the event times are mapped, via a monotonic transformation, to the entire real line. These transformed event times are then written as a function of the covariates. A Gaussian process prior is assumed for the function values. This approach differs from many existing methods of analysing survival data which usually assume some parametric (or semi-parametric) hazard rate. The Cox proportional hazards model (Cox, 1972) is one of the most popular.

In the second route we assume a Cox-type hazard rate. Denoting the d-dimensional vector of covariates as x, the Cox-type hazard rate is $\pi\{\tau \mid f(x)\} = \lambda_0(\tau)\exp\{-f(x)\}$ where $\lambda_0(\tau)$ is the base hazard rate, $\Lambda_0(\tau) = \int_0^\tau \lambda_0(s)\mathrm{d}s$ and $f(x)$ is some function of the covariates. A Gaussian process prior will be assumed for these function values. In Cox's original model $f(x) = \beta^{\mathrm{T}}x$ where β is a d-dimensional vector of regression weights. Such Cox-type models have been

studied by Savitsky et al. (2011) and Joensuu et al. (2012) but have not been applied to interval censored observations.

The first approach (which we refer to as the event time model to distinguish it from the Cox-type model) is a more direct approach since it connects the two quantities which we observe, namely the event times and covariates. We can regard the specification of the hazard rate in Cox-type approaches as an intermediate step in establishing this connection, since the hazard rate is first constructed from the covariates and subsequently used to construct the probability density over event times. Furthermore, in the Cox-type model it is assumed that the time dependence and the covariate dependence of the hazard rate factorizes, but in general this won't always be valid. In the event time model we avoid this assumption on the hazard rate and furthermore we don't have to specify the base hazard rate. The disadvantage however, is that a transformation of the event times must be specified instead.

In the case of either model a full likelihood function can be constructed that can easily incorporate any type of censored or truncated observations. In particular, we apply both models to the case of interval censored data. We numerically maximise the likelihood with respect to the latent function values.

There are three broad families of existing models for analysing interval censored data. Non-parametric estimators based on survival functions that are constant within disjoint intervals have been proposed by Peto (1973) and Turnbull (1976). Secondly, parametric models assume a specific parametric event time density. Popular choices are the Weibull, exponential or log-Gaussian densities for example. The advantage of parametric models is that expressions for the survival function can be obtained in closed form and hence the exact likelihood can be constructed for right, left or interval censored observations. Covariate effects can be included via a link function which specifies that some parameter of the probability density is a function of the covariates. Numerical methods can be used to infer unknown parameter values. See Lindsey (1998) for a discussion and comparison of several parametric models. Odell et al. (1992), Rabinowitz et al. (1995) and Komárek & Lesaffre (2009) consider Weibull accelerated failure time models. Sparling et al. (2006) present a family of parametric models that can handle time dependent covariates.

Finally, there are semi-parametric models, of which most are adaptations of the Cox proportional hazards model. The partial likelihood argument used by Cox cannot be used in the presence of interval censoring. However, the full likelihood can be written in terms of the event time density and survival functions and this can be numerically optimised with respect to any model parameters (Finkelstein, 1986). Markov Chain Monte Carlo methods have been used by Sinha et al. (1999) in a Bayesian discretised Cox model and by Satten (1996) in a proportional hazards model. The EM algorithm has been used by Goggins et al. (1998) and Goetghebeur & Ryan (2000) to infer parameters in proportional hazards models. Several authors use smoothing techniques to model the base hazard rate (Betensky et al., 2002) or the event time density (Zhang & Davidian, 2008). Kooperberg & Clarkson (1997) and Zhang et al. (2010) used splines to model a smooth hazard rate. Another strategy is to impute the event times (Law & Brookmeyer, 1992) by taking the midpoint or the end of the interval for instance. See Pan (2000) for an example.

Before we continue we first consider some notation. Let $p(\tau)$ denote the probability density over the event times $\tau \geq 0$. The survival function is $S(\tau) = \int_\tau^\infty p(s)\mathrm{d}s$ and the hazard rate is $\pi(\tau) = p(\tau)/S(\tau)$. In what follows In this paper we will only consider the case of a single risk with independent censoring. Under that assumption, if we assume a specific form for any one of these quantities then the remaining two quantities are uniquely determined.

2. NON-LINEAR TRANSFORMATION MODELS

Before we explain the Gaussian process models in more detail it is helpful to define a general transformation between covariates and event times:

$$\phi(\tau_i) = f(x_i) + \xi_i, \quad (i = 1, \ldots, n), \tag{1}$$

where ϕ is a monotonically increasing transformation of the event times, $f(x_i)$ is some function of the covariates, ξ_i is a noise random variable with a probability density function denoted by p_ξ, and n is the number of individuals in our dataset. Several existing models, including both Gaussian process models, can be derived as special cases of (1).

Linear transformation models assume ϕ is unspecified and $f(x) = \beta^{\mathrm{T}} x$. Various procedures for estimating the regression parameters in such models have been proposed by Cheng et al. (1995), Fine et al. (1998) and Chen et al. (2002). Recently Lu & Li (2008) considered the case where $f(x)$ is an unspecified smooth function and proposed a boosting estimation method based on the marginal likelihood.

If we choose $p_\xi(s) = \exp(s + e^s)$ and $\phi(\tau) = \log\{\Lambda_0(\tau)\}$ we recover the Cox-type model. To see this we consider the event time to be a transformation of the noise random variable and derive the event time density

$$\begin{aligned} p(\tau) &= p_\xi[\log\{\Lambda_0(\tau)\} - f(x)] \frac{\mathrm{d}}{\mathrm{d}\tau}[\log\{\Lambda_0(\tau)\} - f(x)] \\ &= \lambda_0(\tau) e^{-f(x)} \exp\{-\Lambda_0(\tau) e^{-f(x)}\}. \end{aligned} \tag{2}$$

We can readily verify that this corresponds to a Cox-type hazard rate by writing the survival function in terms of the hazard rate and equating this to the expression for the survival function in terms of the event time density:

$$1 - \int_0^\tau p(s) \mathrm{d}s = \exp\{-\Lambda_0(\tau) e^{-f(x)}\}.$$

By differentiating this expression with respect to time we obtain (2). When $f(x) = -\beta^{\mathrm{T}} x$ we recover Cox's proportional hazards model. Frailty models can be retrieved by assuming $f(x) = -\beta^{\mathrm{T}} x + w$ where w is a frailty term. Generalized additive models assume $f(x) = \beta^{\mathrm{T}} x + \sum_{\mu=1}^{d} g_\mu(x_\mu)$ where g_μ are non-linear functions of the covariates (Fahrmeir & Kneib, 2011). See Martino et al. (2011) and Vanhatalo et al. (2013) for recent implementations of such models. Alternatively, a Gaussian process prior can be assumed for f as shown by Savitsky et al. (2011) and Joensuu et al. (2012). Viewed in this order these models seek to accommodate increasingly complicated covariate effects through more flexible and sophisticated functions of the covariates.

For completeness we note that accelerated failure time models can be recovered by assuming $\phi(\tau) = \log(\tau)$ and $f(x) = \beta^{\mathrm{T}} x$. This implies a log linear model and by choosing different noise distributions a wide range of parametric models can be recovered. See Klein & Moeschberger (2003, Chapter 12) for an overview.

3. GAUSSIAN PROCESS MODELS FOR TIME TO EVENT DATA

3·1. *Gaussian process models*

We will now explain the Gaussian process models for survival analysis in more detail. In general we will observe a pair of event times, I_i, for each individual. For interval censoring $I_i = \{\tau_i^l, \tau_i^u\}$ where τ_i^l and τ_i^u are the lower and upper bounds of an interval within which the

event occurred. For left or right censoring we record the time of censoring and for non-censored individuals the event time is recorded. We also record an indicator variable, Δ_i, which labels which type of censoring each observation corresponds to. If we were to make observations that were both censored and truncated then we would observe more than two times. We will not consider that case explicitly here for brevity although it is straightforward to do so.

In a Gaussian process model we assume the output variables are somehow related to a latent function of the covariates. A Gaussian process model consists of three elements. Firstly, we give the latent function a Gaussian process prior, $f \sim \mathrm{GP}(\eta, k)$, where η is the mean function and k is the kernel function, such that $\mathrm{mean}\{f(x)\} = \eta(x)$ and $\mathrm{cov}\{f(x_i), f(x_j)\} = k(x_i, x_j)$.

Secondly, we have the individual data likelihood terms $\psi(I_i, \Delta_i \mid F_i)$ which will depend on what kind of model has been assumed. Finally, we use Bayes' theorem to obtain the posterior density over the latent function values

$$p(F \mid X, D) = \frac{p(D \mid F)p(F \mid X)}{p(D \mid X)}, \tag{3}$$

where $D = \{(I_1, \Delta_1), \ldots, (I_n, \Delta_n)\}$, $X = (x_1, \ldots, x_n)$, and F is the n-dimensional vector of latent function values such that $F_i = f(x_i)$, and $p(F \mid X)$ is the Gaussian process prior. The data likelihood factorizes over samples so that $p(D \mid F) = \prod_{i=1}^{n} \psi(I_i, \Delta_i \mid F_i)$.

3·2. *Construction of the data likelihood*

Taking the negative log of (3) we get

$$L(F) = -\frac{1}{n}\sum_{i=1}^{n} \log \psi(I_i, \Delta_i \mid F_i) - \frac{1}{n}\log p(F \mid X). \tag{4}$$

We now drop F_i from our notation for brevity. For non-censored individuals $\psi(I_i, \Delta_i) = p(\tau_i)$ which is simply the probability density evaluated at τ_i. Right censored individuals contribute with $\psi(I_i, \Delta_i) = S(\tau_i)$. This gives the probability that the event occurs at some time after τ_i, the time of censoring. Similarly left censored and interval censored individuals contribute with $1 - S(\tau_i)$ and $S(\tau_i^l) - S(\tau_i^u)$ respectively.

Left truncation occurs when only individuals who experience the primary event after a certain time τ_i^a are included in the cohort. The probability of an event occurring at a subsequent time τ_i^b is now conditional on survival until at least τ_i^a. Such an individual will contribute to the likelihood with $\psi(I_i, \Delta_i) = p(\tau_i^b)/S(\tau_i^a)$. Similarly right and interval truncated individuals contribute with $\psi(I_i, \Delta_i) = p(\tau_i^b)/\{1 - S(\tau_i^c)\}$ and $\psi(I_i, \Delta_i) = p(\tau_i^b)/\{S(\tau_i^a) - S(\tau_i^c)\}$ respectively where τ_i^c is the time of right truncation. This can be easily extended to combine any type of censoring with any type of truncation.

3·3. *The event time Gaussian process model*

There are two main challenges to applying standard Gaussian process regression to survival data. The first is to deal with the fact that the outputs are non-negative. This is relatively easy to do with an appropriate transformation of the event times to the entire real line. Secondly, we must deal with the different types of censoring and truncation.

The first step is to transform the event times to $t = \phi(\tau) = \gamma \log(e^{\tau/\gamma} - 1)$, such that t can take any real value. We have chosen this transformation because when $\tau \gg \gamma$ then $t \approx \tau$ and one can tune the value of γ such that $\gamma < \min(\tau_i)$ which results in an effectively linear mapping. We assume a model of the form

$$t_i = f(x_i) + \xi_i, \quad \xi_i \sim N(0, \beta^2) \quad (i = 1, \ldots, n), \tag{5}$$

from which it follows that $t_i \sim N(f(x_i), \beta^2)$. The survival function is given by $S(t_i) = 1 - \Phi(t_i)$ where $\Phi(t_i)$ is the cumulative Gaussian distribution. Note that (5) is a special case of the more general transformation (1). The event time density and survival functions can be simply inserted into (4) to obtain the log likelihood for the event time model (after the event times are transformed).

A similar approach was taken by Chu & Ghahramani (2005) to develop a Gaussian process method for ordinal regression. Ordinal regression is used when the outcomes are assigned to discrete categories which are ranked (exam grades are an example). Mathematically, this is identical to the problem of interval censored event times since the outcome is known only to lie within a certain interval.

The transformation of output variables in Gaussian process regression has been explored by Snelson et al. (2004). They examine a variety of parameterised monotone transformations and regard any transformation parameters as hyperparameters to learn during training. Their procedure infers the most appropriate transformation such that the transformed outputs can be modelled using a Gaussian process. It may be useful to apply this method in future work.

3·4. *The Cox-type Gaussian process model*

In the Cox-type model the event time density is given by (2). The corresponding survival function is $S(\tau_i) = \exp\{-\Lambda_0(\tau_i) f(x_i)\}$. These expressions can also be inserted into (4) to obtain

$$L_C(F) = -\frac{1}{n} \sum_{i:\Delta_i=1} \{\log \lambda_0(\tau_i) + F_i\} + \frac{1}{n} \sum_{i=1}^{n} \Lambda_0(\tau_i) e^{F_i} - \frac{1}{n} \log p(F \mid X). \tag{6}$$

In the Cox-type model none of the event times are transformed. In Section 6 we implement this model with a Weibull hazard rate $\lambda_0(\tau) = \nu \tau^{\nu-1}$ where $\nu > 0$ is optimized as a hyperparameter.

4. Inference and Prediction

4·1. *Inference of latent parameters and hyperparameters*

To determine the optimal latent function values we solve $\hat{F} = \min_F L(F)$ by numerically minimising (4) with a gradient based optimizer. First order partial derivatives can be found in Appendix A. Let θ be the vector of hyperparameters which include any kernel parameters and parameters such as β^2 or ν. To estimate the values of these hyperparameters we require the marginal likelihood which is given by $p(D \mid X) = \int p(D \mid F) p(F \mid X) \mathrm{d}F$. In general, this integral is analytically and numerically intractable. Instead we construct a Laplace approximation of the marginal likelihood. This is done by expanding the log likelihood to second order and then integrating over F. For the Cox-type model for example,

$$\begin{aligned} q(D \mid X) &= \int e^{-nL(\hat{F}) - (F-\hat{F})\cdot(U+K^{-1})(F-\hat{F})/2} \mathrm{d}F \\ &= p(D \mid \hat{F}) p(\hat{F} \mid X) (2\pi)^{n/2} \det\{(U + K^{-1})^{-1}\}^{1/2}, \end{aligned} \tag{7}$$

where $n^{-1}(U + K^{-1})_{ij} = \partial^2 L_C(F) / \partial F_j \partial F_i$ is the matrix of second order partial derivatives given in Appendix A. The matrix $K_{ij} = k(x_i, x_j)$ is the covariance matrix from the Gaussian process prior. For the event time model we simply use the matrix W instead of U (see Appendix A). Optimal hyperparameters are determined by numerically solving $\hat{\theta} = \min_\theta \{-n^{-1} \log q(D \mid X)\}$. This approach is similar to Gaussian process classification (Rasmussen & Williams, 2006, Chapter 3) where the function values are unobserved and treated as latent variables.

4·2. *Predictions, hazard rates and survival functions*

Given a new individual with covariates $x_\star$ we want to predict a corresponding event time $\tau_\star$. In standard Gaussian process regression $f(x_\star) \sim N(\mu, \kappa)$ with $\mu = k_\star^{\mathrm{T}} K^{-1} F$ and $\kappa = k(x_\star, x_\star) - k_\star^{\mathrm{T}} K^{-1} k_\star$ where $k_\star$ is the n-dimensional vector defined by $(k_\star)_i = k(x_i, x_\star)$.

In the event time model we use the mode of the posterior such that $f(x_\star) \sim N(\hat{\mu}, \hat{\kappa})$ with $\hat{\mu} = k_\star^{\mathrm{T}} K^{-1} \hat{F}$ and $\hat{\kappa} = k(x_\star, x_\star) - k_\star^{\mathrm{T}} (K + W^{-1})^{-1} k_\star$. The variance is a combination of uncertainty due to conditioning on $\hat{F}$ and the uncertainty in the value of $\hat{F}$ itself (Rasmussen & Williams, 2006, Section 3.4.2). The predictive distribution for transformed event times is $t_\star \sim N(\hat{\mu}, \hat{\kappa} + \beta^2)$. Finally, we use the transformation ϕ to derive the predictive density for the actual event time:

$$g(\tau_\star \mid X, x_\star, D) = \frac{\exp\left[-\frac{1}{2(\hat{\kappa}+\beta^2)}\{\gamma \log(e^{\tau_\star/\gamma} - 1) - \hat{\mu}\}^2\right]}{\{2\pi(\hat{\kappa} + \beta^2)\}^{1/2}} \frac{e^{\tau_\star/\gamma}}{e^{\tau_\star/\gamma} - 1}. \tag{8}$$

Predictions are made by numerically computing the mean and variance of (8). Due to the transformation ϕ any predictions that are negative or close to zero will be 'squashed' into the positive half of the real line. As noted in the introduction the survival function and hazard rate can be computed once the event time density is known. Some numerical issues are discussed in Appendix B.

In the case of the Cox-type model predictions are made by substituting $\mathrm{mode}(F_\star) = \hat{\mu}$ for $f(x_\star)$ into (2) and using this to numerically compute the mean and variance of $\tau_\star$. Ideally we would multiply the event time density by the predictive density over $F_\star$ and integrate over $F_\star$. However this results in a predictive density over $\tau_\star$ that is not defined. Consequently, by using only the mode of $F_\star$ we underestimate the variance of $\tau_\star$.

5. COMPARISON TO THE COX PROPORTIONAL HAZARDS MODEL

In the results section we will compare the performance of both Gaussian process models to Cox's original proportional hazards model. Cox's model is obtained from (2) when $f(x_i) = -\beta^{\mathrm{T}} x_i$. We use Breslow's estimator for the base hazard rate, originally proposed in the discussion at the end of Cox (1972). This can be written as a maximum likelihood estimator of the base hazard rate (Coolen & Holmberg, 2014, Section 8.1):

$$\lambda_0(\tau) = \frac{\sum_{i=1}^n \delta_{1,\Delta_i} \delta(\tau - \tau_i)}{\sum_{j=1}^n \exp(\beta^{\mathrm{T}} x_j) \theta(\tau_j - \tau)}. \tag{9}$$

This can be inserted into (2) to obtain the event time density corresponding to Cox's model. However, this density is not normalised as can be seen from

$$Z(x_\star) = \int_0^\infty \lambda_0(s) e^{\beta^{\mathrm{T}} x_\star} \exp\{-\Lambda_0(s) e^{\beta^{\mathrm{T}} x_\star}\} \mathrm{d}s = 1 - \exp\{-\Lambda_0(s) e^{\beta^{\mathrm{T}} x_\star}\}\Big|_{s=\infty}.$$

The problem arises since $\Lambda_0(\tau)$ does not diverge to infinity in the limit that τ becomes infinitely large. This is reflected in the fact that any survival curve generated using Cox's model with Breslow's estimator has a non-zero value for infinitely large times. This implies that there is non-zero probability of the event never occurring which is consistent with $Z(x_\star) < 1$.

We normalise the event time density by dividing by $Z(x_\star)$ and use this to compute the mean (and variance) of $\tau_\star$ which are then used for making predictions:

$$\langle \tau_\star \rangle = \frac{1}{Z(x_\star)} \sum_{i:\Delta_i=1} \tau_i \frac{e^{\beta^{\mathrm{T}} x_\star}}{\sum_j e^{\beta^{\mathrm{T}} x_j} \theta(\tau_j - \tau_i)} \exp\left\{ -e^{\beta^{\mathrm{T}} x_\star} \sum_{k=1}^{n} \frac{\delta_{1,\Delta_k} \theta(\tau_i - \tau_k)}{\sum_j e^{\beta^{\mathrm{T}} x_j} \theta(\tau_j - \tau_k)} \right\}. \quad (10)$$

6. RESULTS

Presented in Fig. 1 are results from simulated data with a non-monotonic relationship between the event times and covariates. There is one covariate x and $n = 25$ individuals, of which twelve are censored. In all figures the vertical axis represents the untransformed event time. An end of trial cut-off is imposed at six years and is represented by the dashed line. Any individuals alive at that time are considered right censored. Panel (a) shows the observed data. The black line is the true function which was drawn from a Gaussian process prior with a squared exponential kernel $k(x_i, x_j) = \sigma \exp\{-(2l^2)^{-1}(x_i - x_j)^{\mathrm{T}}(x_i - x_j)\}$ with hyperparameters set to $(\eta, \beta, \sigma, l) = (5, 0.2, 3, 0.7)$. Results from applying the normalized Cox proportional hazards model are shown in panel (b). This is described above in Section 5. We found $\beta = 0.49$. In panel (c) are results from the event time model. Note the $\hat{F}_i$ are plotted and not τ_i. For censored individuals the $\hat{F}_i$ provide an estimate of when the event would have been reported in the absence of censoring or noise. In particular, event times can be estimated several years after the trial ended (note the uncertainty is greatest in this region). Hyperparameters were found to be $(\eta, \beta, \sigma, l) = (5.82, 0.32, 2.59, 0.64)$. In panel (d) the Cox-type model is applied to the data with a squared exponential kernel. Optimal hyperparameters were found to be $(\eta, \nu, \sigma, l) = (-28.4, 16.2, 34, 0.68)$. Note that the variance is underestimated since the the mode of $F_\star$ is used. In panels (e) and (f) we converted non-censored individuals into interval censored individuals by generating a random one year interval for each individual. The event time model is used in (e). Note that $\hat{F}_i$ are plotted and not the observed event times. Optimal hyperparameters were found to be $(\eta, \beta, \sigma, l) = (5.67, 0.14, 3.34, 0.57)$. In (f) are results from the Cox-type model. Optimal hyperparameters were found to be $(\eta, \nu, \sigma, l) = (-45.5, 27.2, 64.7, 0.47)$.

7. DISCUSSION

We have formulated two different Gaussian process models for the analysis of censored survival data. The first model expresses the transformed event time as a function of the covariates whereas the second model is defined by assuming a Cox-type hazard rate. One drawback of the Cox-type model is that the assumed independence of time and covariate effects on the hazard rate may not always be valid. When it is valid however, we can interpret the latent function values as either amplifying or diminishing the base hazard rate for each individual. Thus, negative function values correspond to longer survival times. But when the assumption is invalid then this interpretation may be misleading. In contrast, the role of the latent function values is always clear in the event time model since they can be seen as the event times in the absence of noise or censoring. A second drawback of the Cox-type model is that we are unable to fully quantify the uncertainty of our predictions which leads to an underestimation of the predictive variance.

The first model takes a more direct approach by connecting the two quantities we observe (event times and covariates) directly. In the practical analysis of biomedical survival data, often the most relevant issues are making predictions for new patients and assessing the most relevant covariates. In both cases the most relevant quantity is not the hazard rate, but rather the connection between covariates and event times.

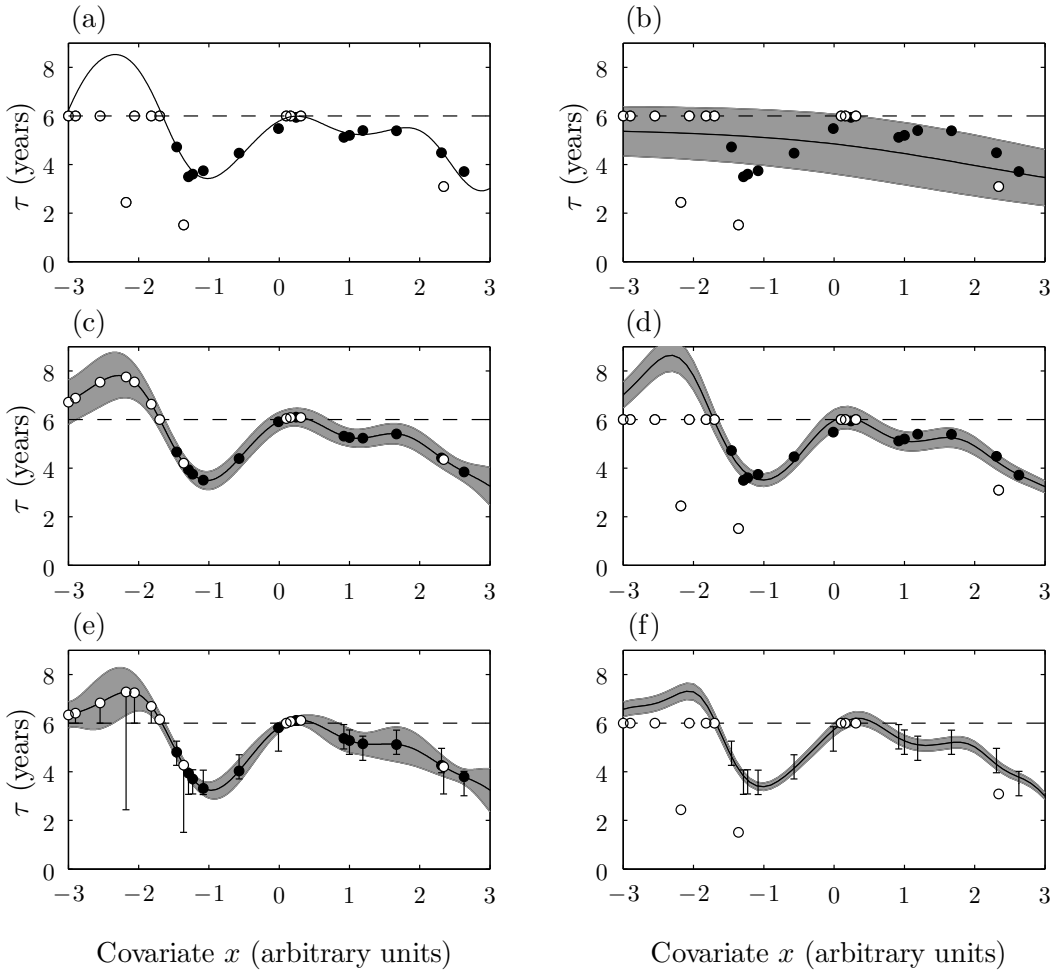

Fig. 1: In panels (a), (b), (d) and (f) the white and black dots correspond to observed right censoring times and non-censored event times respectively. In all figures the dashed line represents an end of trial cutoff time. The observed data are plotted in panel (a). These were drawn from the true function which is represented by the black line. Results from the normalized Cox model are shown in (b). The black line is the mean prediction and the grey area is plus and minus one standard deviation. In panels (c) and (d) are predictions generated using the event time model and the Cox-type model respectively. The variance is underestimated in the Cox-type model since the uncertainty in $\hat{F}$ is not taken into account. In (c) and (e) the white and black dots are $\hat{F}$ for right censored and non-censored individuals respectively. In the bottom two panels, (e) and (f), are results from the event time model and Cox-type models on a combination of interval and right censored data. The 'error bars' denote randomly generated one year censoring intervals. In (e) right censored individuals have a half 'error bar' connecting $\hat{F}_i$ to the time of censoring.

In general, it is natural for a Bayesian analysis of survival data to avoid any partial likelihood arguments and infer the relevant parameters from full likelihoods combined with priors. This offers an intuitive and systematic way to construct the posterior likelihood. The advantage of this in our case is that censored and truncated individuals (and any combination thereof) can be easily incorporated into the likelihood.

Gaussian process models possess several attractive features for the analysis of survival data. Firstly, by specifying and combining different kernels in the Gaussian process prior we can infer a wide range of non-linear relationships between covariates and event times. Secondly, the use of kernel functions means that high dimensional covariates can be studied. The risk of overfitting can be kept to a minimum provided we exclude variable selection hyperparameters. Thirdly, estimates of event times for censored individuals can be extracted from the data. Finally, any type of censoring or truncation can be easily dealt with.

In future work it may be interesting to explore different choices for the noise distribution in our model. In the case of biomedical data it is interesting to ask what the noise represents. Does it represent inherent stochasticity in biological systems or delays in the reporting of events? In the latter case the latent function values can be interpreted as noise-free event times and the noise (with semi-infinite support) represents a waiting time until the event is recorded. For example, it may take some time for symptoms to manifest themselves. In the former case noise with infinite support is more appropriate. In real data both sources of noise are likely to be present and so it would be interesting to implement asymmetric noise distributions that place more density on positive noise values. We also plan to study the case of multiple risks which may or may not be correlated. A potential route would be to use multi-output Gaussian process priors to capture any correlation between the different risks.

ACKNOWLEDGEMENT

This work was funded by the European Union FP7 Imagint Project, grant number 259881.

A. PARTIAL DERIVATIVES

In Section 4·1 we required the first order partial derivatives of the log likelihood. Here we consider non-censored individuals and right censored individuals. For the event time model the first order derivatives are

$$\frac{\partial}{\partial F_i}L(F) = -\frac{1}{n}\sum_{k=1}^{n}\frac{\partial}{\partial F_i}\log\psi(I_k,\Delta_k \mid F_k) + \frac{1}{n}\{K^{-1}(F-\eta)\}_i,$$

where for non-censored individuals $\psi(I_k,\Delta_k \mid F_k) = N(F_k,\beta^2)$ and

$$\frac{\partial}{\partial F_i}\log\psi(I_k,\Delta_k \mid F_k) = \delta_{ik}\beta^{-2}(t_k - F_k).$$

Individuals who were right censored have $\psi(I_k,\Delta_k \mid F_k) = S(t_k \mid F_k)$

$$\frac{\partial}{\partial F_i}\psi(I_k,\Delta_k \mid F_k) = \delta_{ik}\frac{1}{S(t_k \mid F_k)}\frac{1}{(2\pi\beta^2)^{1/2}}e^{-\frac{1}{2\beta^2}(t_k-F_k)^2}.$$

Second order partial derivatives are also required in Section 4·1 to construct a Laplace approximation of the posterior. These are

$$\frac{\partial^2}{\partial F_j\partial F_i}L(F) = -\frac{1}{n}\delta_{ij}\delta_{ik}\sum_{k=1}\frac{\partial^2}{\partial F_j\partial F_i}\log\psi(I_k,\Delta_k \mid F_k) + \frac{1}{n}K^{-1}_{ij} = \frac{1}{n}(W+K^{-1})_{ij},$$

where the diagonal matrix W is defined by $W_{ii} = -\partial^2/\partial F_i^2\log\psi(I_i,\Delta_i \mid F_i)$. For non-censored individuals $W_{ii} = \beta^{-2}$. For right censored individuals

$$W_{ii} = \left\{\frac{1}{S(t_i \mid F_i)}\frac{1}{(2\pi\beta^2)^{1/2}}e^{-\frac{1}{2\beta^2}(t_i-F_i)^2}\right\}^2 - \frac{(t_i-F_i)}{\beta^2}\left\{\frac{1}{S(t_i \mid F_i)}\frac{1}{(2\pi\beta^2)^{1/2}}e^{-\frac{1}{2\beta^2}(t_i-F_i)^2}\right\}.$$

Partial derivatives for interval censored individuals are calculated similarly. See the Appendix B below for numerical details. Partial derivatives of the Cox-type likelihood (6) are

$$\frac{\partial}{\partial F_i} L_C(F) = -\frac{1}{n}\delta_{1,\Delta_i} + \frac{1}{n}\Lambda_0(\tau_i)e^{F_i} + \frac{1}{n}(K^{-1}F)_i$$
$$\frac{\partial^2}{\partial F_j \partial F_i} L_C(F) = \frac{1}{n}\delta_{ij}\Lambda_0(\tau_i)e^{F_i} + \frac{1}{n}K^{-1}_{ij} = \frac{1}{n}(U + K^{-1})_{ij}.$$

B. NUMERICAL CONSIDERATIONS

To compute hazard rates and partial derivatives for the event time model we need to numerically evaluate $A(t) = p(t)/S(t)$ where $p(t)$ is a normal density with variance β^2 and $S(t)$ is the corresponding survival function. The denominator can be written in terms of the complementary error function as $S(t) = \text{erfc}(h)/2$ where $h = t/(2\beta^2)^{1/2}$. Then $A(t) = C\exp(-h^2)/\text{erfc}(h)$ where $C = 2/(2\pi\beta^2)^{1/2}$. The quantity $A(t)$ becomes numerically unstable for large h since the numerator and denominator both tend towards zero. For $h \gg 0$ we use the asymptotic expansion of the complementary error function (Menzel, 1960):

$$\text{erfc}(h) = \frac{e^{-h^2}}{h\sqrt{\pi}}\left[1 - \frac{1}{2h^2} + \frac{2}{(2h^2)^2} - \frac{8}{(2h^2)^3} + \cdots\right] \qquad \text{for } h \gg 0.$$

This ensures $A(t)$ can be computed without difficulty. It is also clear that for large times the hazard rate is approximately linear.

There is an interesting symmetry in the fact that for Cox-type models the computation of event time densities such as (2) or (10) can be numerically unstable due to the presence of the double exponential. In particular, when dealing with the likelihood contribution made by an interval censored observation we must compute quantities of the form $\log(e^{-x_1} - e^{-x_2})$ where $x_k = -\Lambda_0(\tau^k)e^f$. If x_1 or x_2 are sufficiently large then this will be evaluated as $\log(0)$ if double precision numbers are used. Instead we write this as

$$\begin{aligned}\log(e^{-x_1} - e^{-x_2}) &= \log\{e^{-x_1}(1 - e^{x_1 - x_2})\} \\ &= \begin{cases} -x_1 + \log(1 - e^{x_1 - x_2}) & \text{when } -h' \le x_1 - x_2 \\ -x_1 - e^{x_1 - x_2} - \frac{1}{2}e^{2(x_1 - x_2)} & \text{when } x_1 - x_2 < -h'. \end{cases}\end{aligned}$$

where h' is some suitably defined cutoff constant. Similar tricks can be used to ensure the partial derivatives are computed robustly.

REFERENCES

BETENSKY, R. A., LINDSEY, J. C., RYAN, L. M. & WAND, M. (2002). A local likelihood proportional hazards model for interval censored data. *Statistics in Medicine* **21**, 263–275.

CHEN, K., JIN, Z. & YING, Z. (2002). Semiparametric analysis of transformation models with censored data. *Biometrika* **89**, 659–668.

CHENG, S., WEI, L. & YING, Z. (1995). Analysis of transformation models with censored data. *Biometrika* **82**, 835–845.

CHU, W. & GHAHRAMANI, Z. (2005). Gaussian processes for ordinal regression. In *Journal of Machine Learning Research*.

COOLEN, A. C. C. & HOLMBERG, L. (2014). *Principles of Survival Analysis (manuscript in preparation)*. Oxford University Press.

COX, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society: Series B (Methodological)* **34**, 187–220.

FAHRMEIR, L. & KNEIB, T. (2011). *Bayesian Smoothing and Regression for Longitudinal, Spatial and Event History Data*. Oxford University Press.

FINE, J., YING, Z. & WEI, L. (1998). On the linear transformation model for censored data. *Biometrika* **85**, 980–986.

FINKELSTEIN, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics* , 845–854.

GOETGHEBEUR, E. & RYAN, L. (2000). Semiparametric regression analysis of interval-censored data. *Biometrics* **56**, 1139–1144.

GOGGINS, W. B., FINKELSTEIN, D. M., SCHOENFELD, D. A. & ZASLAVSKY, A. M. (1998). A markov chain monte carlo em algorithm for analyzing interval-censored data under the cox proportional hazards model. *Biometrics* , 1498–1507.

JOENSUU, H., VEHTARI, A., RIIHIMÄKI, J., NISHIDA, T., STEIGEN, S. E., BRABEC, P., PLANK, L., NILSSON, B., CIRILLI, C., BRACONI, C., BORDONI, A., MAGNUSSON, M. K., LINKE, Z., SUFLIARSKY, J., FEDERICO, M., JONASSON, J. G., TOS, A. P. D. & RUTKOWSKI, P. (2012). Risk of recurrence of gastrointestinal stromal tumour after surgery: an analysis of pooled population-based cohorts. *The Lancet Oncology* **13**, 265–274.

KLEIN, J. P. & MOESCHBERGER, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer Science+Buisness Media, LLC.

KOMÁREK, A. & LESAFFRE, E. (2009). The regression analysis of correlated interval-censored data illustration using accelerated failure time models with flexible distributional assumptions. *Statistical Modelling* **9**, 299–319.

KOOPERBERG, C. & CLARKSON, D. B. (1997). Hazard regression with interval-censored data. *Biometrics* , 1485–1494.

LAW, C. G. & BROOKMEYER, R. (1992). Effects of mid-point imputation on the analysis of doubly censored data. *Statistics in medicine* **11**, 1569–1578.

LINDSEY, J. (1998). A study of interval censoring in parametric regression models. *Lifetime Data Analysis* **4**, 329–354.

LU, W. & LI, L. (2008). Boosting method for nonlinear transformation models with censored survival data. *Biostatistics* **9**, 658–667.

MARTINO, S., AKERKAR, R. & RUE, H. (2011). Approximate bayesian inference for survival models. *Scandinavian Journal of Statistics* **38**, 514 – 528.

MENZEL, D. H. (1960). *Fundamental Formulas of Physics*, vol. one. Dover Publications, Inc. New York.

ODELL, P. M., ANDERSON, K. M. & D'AGOSTINO, R. B. (1992). Maximum likelihood estimation for interval-censored data using a weibull-based accelerated failure time model. *Biometrics* , 951–959.

PAN, W. (2000). A multiple imputation approach to cox regression with interval-censored data. *Biometrics* **56**, 199–203.

PETO, R. (1973). Experimental survival curves for interval-censored data. *Applied Statistics* , 86–91.

RABINOWITZ, D., TSIATIS, A. & ARAGON, J. (1995). Regression with interval-censored data. *Biometrika* **82**, 501–513.

RASMUSSEN, C. E. & WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA.

SATTEN, G. A. (1996). Rank-based inference in the proportional hazards model for interval censored data. *Biometrika* **83**, 355–370.

SAVITSKY, T., VANNUCCI, M. & SHA, N. (2011). Variable selection for nonparametric gaussian process priors: Models and computational strategies. *Statistical Science* **26**, 130–149.

SINHA, D., CHEN, M.-H. & GHOSH, S. K. (1999). Bayesian analysis and model selection for interval-censored survival data. *Biometrics* **55**, 585–590.

SNELSON, E., RASMUSSEN, C. E. & GHAHRAMANI, Z. (2004). Warped gaussian processes. *Advances in neural information processing systems* **16**, 337–344.

SPARLING, Y. H., YOUNES, N., LACHIN, J. M. & BAUTISTA, O. M. (2006). Parametric survival models for interval-censored data with time-dependent covariates. *Biostatistics* **7**, 599–614.

TURNBULL, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)* , 290–295.

VANHATALO, J., RIIHIMÄKI, J., HARTIKAINEN, J., JYLÄNKI, P., TOLVANEN, V. & VEHTARI, A. (2013). Bayesian modeling with gaussian processes using the gpstuff toolbox. arXiv:1206.5754v4 [stat.ML].

ZHANG, M. & DAVIDIAN, M. (2008). "smooth" semiparametric regression analysis for arbitrarily censored time-to-event data. *Biometrics* **64**, 567–576.

ZHANG, Y., HUA, L. & HUANG, J. (2010). A spline-based semiparametric maximum likelihood estimation method for the cox model with interval-censored data. *Scandinavian Journal of Statistics* **37**, 338–354.

[*Received April* 2012. *Revised September* 2012]