**Preprint of the statistical appendix for**
Heikki Joensuu, Peter Reichardt, Mikael Eriksson, Kirsten Sundby Hall and Aki Vehtari (2013). Gastrointestinal Stromal Tumor: A Method for Optimizing the Timing of CT Scans in the Follow-up of Cancer Patients. In *Radiology*, in press.

In this supplementary file, we describe in a detail how to apply the Gaussian processes (GP) in nonhomogoneus Poisson process survival analysis with interval censored data. This statistical methodology is applied in the paper "Gastrointestinal stromal tumor: a method for optimizing the timing of CT scans in the follow-up of cancer patients".

For the individual $i$, where $i = 1, \ldots, n$, we have survival time $y_i$ (possibly right or interval censored) with a censoring indicator $\delta_i$, where $\delta_i = 0$ if the $i$th observation is uncensored and $\delta_i = 1$ if the observation is right or interval censored. For interval censored survival time, $y_i$ is known to fall into an interval $[y_{i,\mathrm{lo}}, y_{i,\mathrm{up}}]$. The traditional approach to analyze continuous time-to-event data is to assume the Cox proportional hazards function (Cox, 1972)

$$h_i(t) = h_0(t) \exp(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}), \tag{1}$$

where $h_0$ is the unspecified baseline hazard rate, $\boldsymbol{x}_i$ is the $d \times 1$ vector of covariates for the $i$th patient and $\boldsymbol{\beta}$ is the vector of regression coefficients. The matrix $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^{\mathrm{T}}$ of size $n \times d$ includes all covariate observations.

The Cox model with a linear predictor can be extended to more general form to enable, for example, additive and non-linear effects of covariates (Kneib, 2006; Martino et al., 2011). Previously, we (Joensuu et al., 2012) extended the proportional hazards model by

$$h_i(t) = \exp(\log(h_0(t)) + \eta(\boldsymbol{x}_i)), \tag{2}$$

where the linear predictor was replaced with the nonlinear predictor $\eta$ depending on the covariates $\boldsymbol{x}_i$. Here, to allow the form of the hazard function to depend on the covariates, we use even more generic hazard model

$$h_i(t) = \exp(\eta(t, \boldsymbol{x}_i)). \tag{3}$$

A piecewise log-constant hazard in time (Ibrahim et al., 2001) is assumed by partitioning the time axis into $K$ non-overlapping intervals with equal lengths: $0 = s_0 < s_1 < s_2 < \ldots < s_K$, where $s_K \geq y_i$ for all $i = 1, \ldots, n$. In the interval $k$, where $k = 1, \ldots, K$, hazard is assumed to be constant in time and for the $i$th individual the hazard rate in the $k$th time interval is

$$h_i(t) = \exp(\eta(\tau_k, \boldsymbol{x}_{ik})), \quad t \in (s_{k-1}, s_k], \tag{4}$$

where $\tau_k = (s_k - s_{k-1})/2$ is the mean of $k$th time interval and $\boldsymbol{x}_{ik}$ denotes possibly time varying covariates.

Using the piecewise log-constant assumption for the hazard rate function, the contribution of the possibly right censored $i$th observation $(y_i, \delta_i)$ for the likelihood is(Martino et al., 2011; Ibrahim et al., 2001)

$$l_i = [\exp(\eta_{iK_i})]^{(1-\delta_i)} \exp\left( - \sum_{g=1}^{K_i} (s_k - s_{k-1}) \exp(\eta_{ik}) \right), \tag{5}$$

where $\eta_{ik} = \eta(\tau_k, \boldsymbol{x}_{ik})$. This can be replaced with likelihood of $K_i$ Poisson distributed data points for $i$th person, with means $(s_k - s_{k-1}) \exp(\eta_{ik})$, of which $K_i - 1$ first ones are observed to be 0, and the last one observed to be 0 or 1 according to whether the survival time $t$ is observed or censored. Augmented data thus has $\tilde{n} = \sum_{i=1}^{n} K_i$ datapoints. Interval censored survival times $y_i$ can be imputed by sampling from the interval $[y_{i,\mathrm{lo}}, y_{i,\mathrm{up}}]$ according to conditional density $p(\hat{y}_i|h_i)$.

By assuming a Gaussian process prior (Rasmussen and Williams, 2006, e.g.,) over the unknown function $\eta$, smooth nonlinear effects of continuous covariates are possible, and if there are dependencies between covariates, the GP can model these interactions implicitly. A zero-mean GP prior is set for $\eta$, which results in the zero-mean multivariate Gaussian distribution

$$p(\eta|X) = \mathcal{N}(\mathbf{0}, C(X, X)), \tag{6}$$

where $C(X, X)$ is the $\tilde{n} \times \tilde{n}$ covariance matrix whose elements are given by the covariance function of the GP. The covariance function defines the smoothness and scale properties of the latent function, and we choose a sum of constant and non-stationary neural network covariance function (Williams, 1998)

$$c(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sigma_c + \frac{2}{\pi} \sin^{-1}\left(\frac{2\tilde{\boldsymbol{x}}_i^{\mathrm{T}} \Sigma \tilde{\boldsymbol{x}}_j}{(1 + 2\tilde{\boldsymbol{x}}_i^{\mathrm{T}} \Sigma \tilde{\boldsymbol{x}}_i)(1 + 2\tilde{\boldsymbol{x}}_j^{\mathrm{T}} \Sigma \tilde{\boldsymbol{x}}_j)}\right), \tag{7}$$

where $\sigma_c$ is the constant covariance part, $\tilde{\boldsymbol{x}} = (1, \tau, x_1, \ldots, x_d)^{\mathrm{T}}$ is an input vector augmented with 1 and $\tau$, and $\Sigma = \mathrm{diag}(\sigma_0^2, \sigma_\tau^2, \sigma_1^2, \ldots, \sigma_d^2)$ is a diagonal weight prior, where $\sigma_0^2$ is a variance for the bias parameter controlling the functions offset from the origin and $\sigma_\tau^2, \sigma_1^2, \ldots, \sigma_d^2$ are the variances for the weight parameters. The constant covariance part models the mean hazard level, and the neural network covariance part models the nonlinear function. A neural network covariance function was chosen, since it is suitable for modeling non-stationary, saturating and interaction effects. A half-Gaussian prior with variance 4 was used for the constant covariance $\sigma_c$, and half-$t$ priors with degrees of freedom 4 and variances 100 and 10 were used for $\sigma_0$ and $\sigma_\tau, \sigma_1, \ldots, \sigma_d$, respectively, as recommended by Gelman (2006).

By applying the Bayes theorem, the prior information and likelihood contributions are combined to get posterior distribution of the latent variables and the covariance function parameters. To compute predictions, we integrated over the hyperparameters of the covariance function and the unknown $y_i$ for interval censored observations using Markov chain Monte Carlo sampling. The covariance function parameters were sampled using hyperrectangle multivariate slice sampling (Neal, 2003) and the conditional distribution $p(\theta|\boldsymbol{y}, \boldsymbol{x})$ where latent variables $\boldsymbol{\eta}$ were integrated out using a Gaussian approximation. The posterior distribution of the latent variables is approximated by doing a second order Taylor expansion of the logarithm of the posterior around the posterior mode, as presented by Rasmussen and Williams (2006). The unknown $y_i$ for interval censored observations were sampled by first sampling latent values $\eta_{ik}$ from the Gaussian approximation of the conditional distribution $p(\eta|\theta, \boldsymbol{y}, \boldsymbol{x})$, then computing the hazard $h$ given the latent values, and finally sampling $y_i$ from the conditional density $p(y_i|h_i)$ at interval $[y_{i,\mathrm{lo}}, y_{i,\mathrm{up}}]$. Sampling of $\theta$, $\eta$ and $y$ was performed alternately for 1000 iterations. Convergence of the chain was diagnosed using potential scale reduction factor (Brooks and Gelman, 1998). Geyer's initial monotonic sequence estimator (Geyer, 1992) was used to assess that length of the chain is sufficient to get useful efficient sample size. Inference for the model was made using GPstuff toolbox (Vanhatalo et al., 2013).

A transformation $\log(x+1)$ was applied to reduce the skewness of mitotic count. For the posterior inference the time axis was divided in $K = 29$ 3-month intervals. To improve the accuracy of the computations in the CT scan timing optimization, the hazard was predicted in $K = 85$ 1-month intervals. Number of CT scans was varied from 4 to 12 and the last CT scan was fixed at 6 years after the surgery (due to the limited follow-up time in the study, further hazard estimates would be uncertain), and timings were discretized to 3-month accuracy. Criterion to optimise was expected time from the recurrence to the observation given the recurrent free state at the previous CT scan. Optimization was made by computing the expected time from the recurrence to the observation exhaustively for each possible CT scan schedule.

## References


Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7:434–455.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–533.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Statistical Science*, 7(4):473–483.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001). *Bayesian Survival Analysis*. Springer.

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., Massimo, F., Jonasson, J. G., Paolo Dei Tos, A., and Rutkowski, P. (2012). Risk of gastrointestinal stromal tumour recurrence after surgery: an analysis of pooled population-based cohorts. *The Lancet Oncology*, 13(3):265–274.

Kneib, T. (2006). Mixed model-based inference in geoadditive hazard regression for interval censored survival times. *Computational Statistics and Data Analysis*, 51(2):777–792.

Martino, S., Akerkar, R., and Rue, H. (2011). Approximate Bayesian inference for survival models. *Scandinavian Journal of Statistics*, 38(3):514–528.

Neal, R. M. (2003). Slice sampling. *The Annals of Statistics*, 31(3):705–767.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press.

Vanhatalo, J., Riihimäki, J., Hartikainen, J., Jylänki, P., Tolvanen, V., and Vehtari, A. (2013). GPstuff: Bayesian modeling with Gaussian processes. *Journal of Machine Learning Research*, 14:1175–1179.

Williams, C. K. I. (1998). Computation with infinite neural networks. *Neural Computation*, 10(5):1203–1216.