Stockholms universitet

Research Report

Department of Statistics

No. 2018:5

Bayesian Sequential Inference for Dynamic Survival Models

Parfait Munezero

Department of Statistics, Stockholm University, SE-106 91 Stockholm, Sweden

Bayesian Sequential Inference for Dynamic Survival Models

Parfait Munezero

Department of Statistics, Stockholm university

Abstract

Dynamic hazard models are applied to analyze time-varying effects of covariates on the survival time. The state-of-the-art methods for learning parameters in the Bayesian framework are MCMC methods but due to high correlations among the time-varying effect parameters, they converge very slowly. To handle these correlations efficiently, we apply a Sequential Monte Carlo (SMC) method commonly known as Particle Filter (PF). We develop a proposal distribution tailored to the nature of the survival data based on the second order Taylor series expansion of the posterior distribution and the linear Bayes theory. Our PF based sampler is shown to be faster and generates an effective sample size that is more than two orders of magnitude larger than a state-of-the-art MCMC sampler for the same computing time.

Keywords: Hazard function, Linear Bayes, Particle filter, Piecewise exponential, Survival function.

Email address: parfait.munezero@stat.su.se (Parfait Munezero)

1. Introduction

The standard proportional hazards model (Cox, 1972) specifies the hazard function as a product of a baseline hazard which is a function of the time, t, and a relative hazard which is a function of the covariate, $\mathbf{x}$:

$$\lambda(t|\,\mathbf{x}) = \lambda_0\,(t)\exp\left(\mathbf{x}'\beta\right), \tag{1}$$

where β is a vector of parameters to be estimated. Implicit in the above model is the assumption that the ratio of the hazards of any two different values of $\mathbf{x}$ is constant over time. However, in real-life situations the effect of a covariate, and hence the ratio of the hazards, may change over time, especially when the observation period is long. In order to relax this restrictive assumption, the above model can be extended to a generalized hazard model where the parameters are allowed to vary over time:

$$\lambda(t|\mathbf{x}) = \lambda_0(t)\exp\left(\mathbf{x}'\beta(t)\right). \tag{2}$$

This model is a special case of the generalized additive models in Fahrmeir and Kneib (2011) and Hennerfeind et al. (2006) which incorporate nonlinear effects of continuous covariates, spatial effects and frailty terms.

Houwelingen and Putter (2011) express $\beta(t)$ using a set of basis functions, while Hennerfeind et al. (2006) propose a Bayesian approach based on P-splines. Another common and simpler approach of expressing $\beta(t)$ is the semi-parametric piecewise exponential (PE) model, which is a special case of the P-splines with degree zero. The PE parametrizes $\beta(t)$ through a Gaussian random walk process. It has been applied to analyze continuous survival times (Gamerman, 1991; Hemming and Shaw, 2005; Wagner, 2011; Hemming and Shaw, 2002) and discrete survival times (Fahrmeir and Wagenpfeil, 1996; Fahrmeir, 1994).

Although the PE model is simple, it has some drawbacks associated with nonlinearity and correlations among parameters induced by the random walk process. The nonlinearity makes the posterior analytically intractable. Thus, it requires the Metropolis-Hasting within Gibbs algorithm to sample from the posterior distribution (Hemming and Shaw, 2005). On the other hand, correlations among parameters slow down the convergence of the MCMC algorithms (Gamerman, 1998).

Potential solutions to these issues include reparametrization of the random walk process in terms of a sum of Gaussian noise (Hemming and Shaw,

2002) and data augmentation using auxiliary variables (Wagner, 2011). The former manipulates the random walk process that defines the evolution of $\beta(t)$ and applies the Metropolis-Hasting algorithm to sample from the posterior. The latter approximates the observation model based on a data augmentation method that involves approximating the likelihood by a mixture of ten Gaussian distribution. These approaches achieve a better convergence at the expense of costly iterations mostly associated with the MCMC samplers and partly with the model manipulation.

In the present work, we introduce an alternative Sequential Monte Carlo (SMC) approach that, to the best of our knowledge, has not been explored specifically for survival models. SMC sampling methods commonly known as particle filter (PF) algorithms, are generally more suitable for dynamic models in particular nonlinear dynamic models (Doucet et al., 2001). The advantage of SMC algorithms is that they do not require any transformation of the model and they have computational advantages compared to MCMC methods (Del Moral et al., 2006).

PF algorithms require designing a good proposal distribution that is simpler to sample from. To do so, we develop a proposal distribution adapted to the nature of survival data based on the second order Taylor series expansion of the posterior distribution around the mode and the linear Bayes method described in West et al. (1985) and Gamerman (1991). We explore the potential of our particle filter algorithm, which we refer to as the particle filter with linear Bayes tailored proposal (PFLiB), for survival models in a filtering context. PFLiB can be used as basis for developing smoothing algorithms (Briers et al., 2010; Fearnhead et al., 2010), and more advanced algorithms such as particle MCMC (Andrieu et al., 2010; Lindsten et al., 2014) and SMC^2 (Chopin et al., 2013) algorithms, but we leave these to future research.

In the next section we discuss the dynamic survival model specifications; its likelihood and prior. Section 3 introduces briefly the parameter updating procedure. In Section 4, we apply the procedures to two known data sets. In the first application the PFLiB is compared to the auxiliary mixture method of Wagner (2011) on a model with one covariate, and in the second application the PFLiB is applied on a model with several covariates. Concluding remarks and suggestions for future research are provided in the last section.

2. Dynamic survival model specification

Let T denote a random variable representing the survival time, which is the time until a certain event of interest occurs. The survival time is not observed for all individuals in the study: some individuals are lost or the study ends before they experience the event. In survival data analysis, such individuals whose survival time is not known are called censored observations. We denote the censoring time by the random variable C. Thus, the observed survival data for n individuals consists of $t_i = \min(T_i, C_i)$, a censoring indicator d_i ($d_i = 0$ if censored, and $d_i = 1$ if event), and a vector of explanatory variables $\mathbf{x}_i$, with $i = 1, \cdots, n$.

The hazard function describes the instantaneous rate at which the event occurs:

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

The survival function $S(t)$ which represents the proportion of individuals who have not experienced the event at time t is related to the hazard function through the expression

$$S(t) = P\left(T > t\right) = \exp\left(-\int_0^t \lambda(s) ds\right).$$

PE models partition the survival time into consecutive time intervals, $I_j = [\tau_{j-1}, \tau_j)$, $j = 1, \cdots, J$, where $\tau_0 = 0 < \tau_1 <, \cdots, < \tau_J$. The baseline hazard function is assumed to be constant within each I_j; i.e. $\lambda_0(t) = \lambda_{0j}$, for $t \in I_j$ and some constant $\lambda_{0j} > 0$. In this case, the log baseline hazard becomes the intercept and the model (2) can be rewritten as

$$\ln \lambda_j = \mathbf{z}' \theta_j, \quad \text{for } t \in I_j \tag{3}$$

where $\mathbf{z} = (1, \mathbf{x}')'$ is the original covariate vector of length p augmented with a column of 1s, and $\theta_j = (\theta_{0j}, \theta_{1j}, \cdots, \theta_{pj})$ is the new regression parameters vector within the interval I_j, with the intercept $\theta_{0j} = \ln(\lambda_{0j})$. The implication of the relation (3) is that the integral in the expression of the survival function is replaced by a summation, thus the survival function for individuals who survived the previous interval I_{j-1} becomes.

$$S\left(t | t > \tau_{j-1}\right) = \exp\left(-\lambda_j \left(t - \tau_{j-1}\right)\right), \; j = 1, \cdots, J. \tag{4}$$

2.1. Likelihood function

For data with censored observations, the likelihood function can be expressed in terms of the hazard and the survival functions as

$$L\left(t|\lambda,d\right)=\prod_{i=1}^{n}\lambda(t_i)^{d_i}S\left(t_i\right). \tag{5}$$

Gamerman (1991) showed that with the assumption of piecewise constant regression parameters, the likelihood (5) can be factorized across the intervals I_j. That is, the full likelihood is expressed in terms of the product of interval-based likelihood contributions that are computed based on the data available within each interval.

Assuming that n_j individuals are at risk at the beginning of interval I_j, the data available within I_j (that we denote by D_j throughout the remaining sections) include the covariate vector $\mathbf{x}_i$, the exposure time t_{ij} spent in I_j, and the censoring indicator d_{ij} ($d_{ij}=0$ if individual i is censored within interval j, else $d_{ij}=1$). The survival time based on the interval data are defined as follows:

$$t_{ij}=\min\left(t_i-\tau_{j-1},\tau_j-\tau_{j-1}\right).$$

The contribution to the likelihood of all individuals in interval I_j is given by

$$\begin{aligned} L_j\left(t_j|\lambda_j\right) &= \prod_{i=1}^{n_j}\lambda_{ij}^{d_{ij}}\exp\left(-\lambda_{ij}t_{ij}\right) \\ \ln\lambda_{ij} &= \mathbf{z}_i'\theta_j \end{aligned} \tag{6}$$

Alternative likelihood expressions can be obtained via a data augmentation scheme that uses auxiliary survival data (Wagner, 2011) or by using a partial likelihood approach (Sargent, 1997).

2.2. Prior specification

In order to complete the model specification, it is important to define prior processes for the regression parameters. One of the simplest and most applied prior on θ_j is the random walk

$$\theta_j=\theta_{j-1}+\epsilon_j,\quad \epsilon_j\sim N\left(0,U_j\right). \tag{7}$$

This process has been adopted by Wagner (2011), Fahrmeir (1994) and Hemming and Shaw (2002). It is a special case of the more general first order

random walk process of the parameter evolution suggested by Gamerman (1991). Clearly, if U_j is zero then there is no change in the parameters and the dynamic model reduces to the standard proportional hazards model. In many applications U_j is held constant; i.e $U_j = U$. Generally, the covariance matrix U is not known, and hence needs to be estimated. Hemming and Shaw (2002) suggests a diagonal matrix U and sets a lognormal prior on the diagonal components. Sargent (1997) and Wagner (2011) set an inverse gamma prior on the diagonal components of U. Another approach suggested by Gamerman (1998) is to apply an inverse Wishart prior on U. Considering a constant covariance U results in a global smoothing, and may not adapt well to abrupt changes or high fluctuations.

On the other hand, allowing the variance to change over time improves the capacity of the random walk process to adapt locally (Fahrmeir and Kneib, 2011). However, estimating U_j dynamically may not be an easy task. West et al. (1985) proposed a way of bypassing the estimation of U_j by applying a discounting procedure. In this procedure, a discount factor is defined in terms of a parameter $0 < \phi < 1$ that controls the amount of information propagated from one interval to the next. If ϕ is close to zero, then no information is propagated across intervals, and the parameters are updated independently. On the other hand, if ϕ is close to one, all information from the previous interval is propagated to the current interval. Though not pursued in this paper, the discount factor can be allowed to vary dynamically (Das and Dey, 2013).

In the next section, we present the posterior distribution and describe briefly the SMC sampling procedure that will be used to sample from the posterior. All inference is based on the model defined by (6) and the prior processes (7). The parameter updating is achieved by iteratively sampling the posterior distribution from interval to interval.

3. Sequential Monte Carlo sampling methods

The whole vector of the regression parameters for all intervals is updated jointly through the posterior distribution

$$p\left(\theta_{1:J}|D_J\right) \propto p\left(t_1|\theta_1\right) p\left(\theta_1\right) \prod_{j=2}^{J} p\left(t_j|\theta_j\right) p\left(\theta_j|\theta_{j-1}\right), \tag{8}$$

where $\theta_{1:J} = (\theta_1, \cdots, \theta_J)$. In order to sample from the joint posterior, a standard MCMC sampling approach requires choosing an initial block of parameters $(\theta_1^0, \cdots, \theta_J^0)$ and updating each block's component by sampling from the conditional distribution $P\left(\theta_j|\theta_{-j}, D_J\right)$. However, this procedure faces two major drawbacks. The first is the correlations among the parameters in different intervals, which slows down the convergence of the MCMC chain. The second is the impossibility to sample directly from the conditional distribution $P\left(\theta_j|\theta_{-j}, D_J\right)$ since it is not analytically tractable due to the nonlinearity embedded in the likelihood.

In order to improve the performance of MCMC algorithms for dynamic models, different approaches have been suggested. These include the forward filtering backward sampling algorithm (Frühwirth-Schnatter, 1994), and the reparametrization of the parameter evolution process (Gamerman, 1998). For a survey of such algorithms see Migon et al. (2005).

In our present case, we will explore an alternative method based on the SMC commonly known as the particle filter (Van Der Merwe et al., 2001). The SMC procedure updates parameters one at a time sequentially using the data available at the current time point through the following two recursive steps:

$$p\left(\theta_j|D_{j-1}\right) = \int p\left(\theta_j|\theta_{j-1}\right) p\left(\theta_{j-1}|D_{j-1}\right) d\theta_{j-1} \tag{9}$$

$$p\left(\theta_j|\, D_j\right) \propto p\left(y_j|\theta_j\right) p\left(\theta_j|D_{j-1}\right), \tag{10}$$

where $D_j = \left\{\left(t_{ij}, x_i, d_{ij}\right), i = 1, \cdots, n_j\right\}$ represents the data of all individual who are at risk in the j^{th} interval. Starting from a known initial distribution $p\left(\theta_1\right)$ at the time point $j = 1$, parameters are drawn from $p\left(\theta_j|D_{j-1}\right)$, and they are updated according to Equation (10). The process evolves from interval to interval until the last interval is reached. However, this is possible only if one is able to draw samples from $p\left(\theta_j|D_{j-1}\right)$; that is, if the integral in

(9) is tractable. Such analytically defined prior exists only for linear Gaussian dynamic models, for which the resulting filtering distribution $p\left(\theta_j|\, D_j\right)$ is a normal distribution with moments obtained recursively by the Kalman filter procedure (Kalman, 1960). On the other hand, nonlinear models such as (6) do not yield a prior distribution that is tractable analytically and hence one can not draw parameters directly from the posterior distribution. In this situation, either the observational model (6) is linearized through some manipulation of the model such as data augmentation (Wagner, 2011), or through a Gaussian approximation of the observational model using a second order Taylor series expansion. PF algorithms apply the latter approach (Van Der Merwe et al., 2001).

Clearly there is a difference between the joint smooth posterior (8) and the filtering distribution (10) in the sense that the former uses the entire data for updating parameter, but the latter uses only a partition of the data. In the present work we are interested in developing a method for sampling from (10), but one can easily extend our method to sample from (8) by applying algorithms in Briers et al. (2010) or Fearnhead et al. (2010). In the next section, we briefly describe the particle filter.

3.1. Particle filter

PF algorithms approximate the posterior distribution $p(\theta_{j-1}|D_{j-1})$ empirically by a sample of K draws $\left\{\theta_{j-1}^{(1)}, \cdots, \theta_{j-1}^{(K)}\right\}$ (known as particles) associated with corresponding weights $\left\{w_{j-1}^{(1)}, \cdots, w_{j-1}^{(K)}\right\}$,

$$w_j^{(k)} \propto w_{j-1}^{(k)} \frac{p\left(t_j|\theta_j^{(k)}\right) p\left(\theta_j^{(k)}|\theta_{j-1}^{(k)}\right)}{q\left(\theta_j^{(k)}|\theta_{j-1}^{(k)}, D_j\right)}, \tag{11}$$

where $q\left(\theta_j^{(k)}|\theta_{j-1}^{(k)}, D_j\right)$ is the proposal distribution. With the set of particles and their associated weights, one can approximate the online posterior distribution at the j^{th} time interval empirically (Arulampalam et al., 2002):

$$p\left(\theta_j|D_j\right) = \sum_{k=1}^{K} w_j^{(k)} \delta\left(\theta_j - \theta_j^{(k)}\right), \tag{12}$$

where $\delta\left(\cdot\right)$ is the Dirac delta function. Thus, the multidimensional joint posterior problem (8) is broken down into a simplified problem that consists of sampling particles $\left\{\theta_j^{(k)}\right\}_{k=1}^{K}$, and computing their corresponding

weights $\left\{w^{(k)}\right\}_{k=1}^{K}$ recursively. This process in known as the sequential importance sampling (SIS) (Liu and Chen, 1998). A sample of particles $\left\{\theta_j^{(k)}\right\}_{k=1}^{K}$ is drawn from the proposal distribution, and their corresponding weight $\left\{w^{(k)}\right\}_{k=1}^{K}$ are updated according to Equation (11). Then, the online posterior is approximated empirically by the sampled particles as in Equation (12).

Although the SIS provides a simple mechanism of sampling sequentially from the posterior, it has one serious drawback. The support of posterior tends to be dominated by the few particles that have significantly large weights. This degeneracy issue leads to a substantial increase of the variance of the weights over time (Van Der Merwe et al., 2001).

To alleviate this problem, a sampling importance resampling (SIR) algorithm was suggested. This step consists in resampling with replacement the sampled particles $\left\{\theta_j^{(k)}\right\}_{k=1}^{K}$ with probability proportional to their corresponding weights $\left\{w^{(k)}\right\}_{k=1}^{K}$. The effect of the resampling step is that the resampled particles will have equal weights, see Van Der Merwe et al. (2001) for more details.

A variant of the SIR algorithm is the auxiliary sequential importance resampling (ASIR) proposed by Pitt and Shephard (1999). Instead of sampling particles according to their weights, the resampling step is performed by sampling particles proportional to

$$\varphi_j^{(k)} = p\left(t_j|\,\mu_{j-1}^{(k)}\right) w_{j-1,}^{(k)} \tag{13}$$

where $\mu_{j-1}^{(k)}$ represents either a draw from $p\left(\theta_j|\theta_{j-1}\right)$ or the expected value $E\left[\theta_j|\theta_{j-1}^{(k)}\right]$(Arulampalam et al., 2002). Setting $\varphi_j^{(k)} = p\left(t_j|\,\theta_{j-1}^{(k)}\right) w_{j-1}^{(k)}$ results in an adapted filter with all current weights w_j distributed uniformly (Fearnhead et al., 2010), hence the weights are updated as follows:

$$w_j^{(k)} \propto \frac{p\left(t_j|\theta_j^{(k)}\right) p\left(\theta_j^{(k)}|\theta_{j-1}^{(k)}\right) w_{j-1}^{(k)}}{q\left(\theta_j^{(k)}|\theta_{j-1}^{(k)}, D_j\right) \varphi_j^{(k)}}. \tag{14}$$

The efficiency of the particle filter depends on a good proposal distribution design (Van Der Merwe et al., 2001). Therefore, the choice of the proposal distribution is one of the main issues for the PF algorithm.

3.2. Proposal distribution

There are many ways of designing the proposal distribution $q\left(\theta_j|\theta_{j-1}, D_j\right)$. A simple choice is to simply use the transition distribution $p\left(\theta_j|\theta_{j-1}\right)$. This choice has some computational advantages since the computation of the weights can be parallelised easily. However, it is not the optimal distribution. It ignores the information from data $p\left(t_j|\theta_j\right)$ resulting in higher variance of the weights (Van Der Merwe et al., 2001). Common designs of optimal proposal distribution rely on the local linearization (Doucet et al., 2000), or the unscented transformation of the posterior (Van Der Merwe et al., 2001).

To design our proposal distribution, we combine the local linearization of the posterior with the linear Bayes method (West et al., 1985) and Gamerman (1991). To do so, we approximate the true posterior (9) by a normal distribution obtained from a second-order Taylor series expansion of the log posterior around the mode. Observe that from the model in Equation (6), θ_j are linked to the likelihood only through the function $\ln \lambda_{ij} = \mathbf{z}_i'\theta_j$. Therefore, the posterior update of θ_j depends only on $\ln \lambda_{ij}$ which is not an observed quantity. Linear Bayes methods described in Section 2.2 of West et al. (1985) update the posterior of θ_j through the change in mean and variance of $\ln \lambda_{ij}$.

Before proceeding, let us define some new notations that will be used. We denote by $D_{i,j} = (t_{ij}, x_i, d_{ij})$ the data available from the i^{th} individual at risk at the beginning of the j^{th} interval. Further, we treat the log hazard as a random variable $\eta_{ij} = \ln \lambda_{ij} = \mathbf{z}_i'\theta_j$. The contribution of the i^{th} individual to the joint posterior $p\left(\theta_j, \eta_j|\theta_{j-1}, D_j\right)$ is given by the expression (West et al., 1985)

$$\begin{aligned} p\left(\theta_j, \eta_{ij}|\theta_{j-1},\, D_{i,j}\right) &\propto p\left(t_{ij}|\eta_{ij}\right) p\left(\eta_{ij}|\theta_{j-1}, D_{i,j-1}\right) p\left(\theta_j|\eta_{ij}, \theta_{j-1}, D_{i,j-1}\right) \\ &\propto p\left(\eta_{ij}|D_{i,j}\right) p\left(\theta_j|\eta_{ij}, \theta_{j-1}, D_{i,j-1}\right). \end{aligned} \tag{15}$$

Note that the joint posterior $p\left(\theta_j, \eta_{ij}|\theta_{j-1},\, D_{i,j}\right)$ is degenerate because η_{ij} and θ_j are linked through a deterministic relationship. Therefore, a prior on η_{ij} implies a prior on θ_j through the relationship $\eta_{ij} = \mathbf{z}_i'\theta_j$. The aim of defining the joint posterior (15) is to express the proposal distribution $q\left(\theta_j^{(k)}|\theta_{j-1}^{(k)}, D_j\right)$ as the Gaussian approximation of the marginal $p\left(\theta_j|\theta_{j-1}, D_{i,j}\right)$. To do so, we need to approximate the first term in the right hand side of Equation (15) by a Gaussian distribution and apply the linear Bayes theory (West et al., 1985).

Given the structure of the likelihood (6), the original parameter λ_{ij} has a gamma distribution conjugate prior, $Gamma\left(\alpha_{ij}, \psi_{ij}\right)$, and it follows that the marginal posterior of λ_{ij} is $Gamma\left(\alpha_{ij} + d_{ij}, \psi_{ij} + t_{ij}\right)$. Taking into account the Jacobian of the transformation $\eta_{ij} = \ln \lambda_{ij}$,

$$p\left(\eta_{ij}|D_{ij}\right) \propto \exp\left\{\eta_{ij}\left(\alpha_{ij} + d_{ij}\right) - \left(\psi_{ij} + t_{ij}\right)\exp\left\{\eta_{ij}\right\}\right\} \tag{16}$$

The Gaussian approximation of the posterior (16) is obtained from the following approximation

$$p\left(\eta_{ij}|D_{ij}\right) \approx N\left(\left[\frac{\partial \ln p\left(\eta_{ij}|D_{ij}\right)}{\partial \eta_{ij}}\right]_{\eta_{ij}=\hat{\eta}_{ij}}, \left[-\frac{\partial^2 \ln p\left(\eta_{ij}|D_{ij}\right)}{\partial \eta_{ij}^2}\right]^{-1}_{\eta_{ij}=\hat{\eta}_{ij}}\right), \tag{17}$$

where $\hat{\eta}_{ij}$ is the mode. In our particular case, the first and second derivatives of the log of the posterior (16) are given by $\frac{\partial \ln p(\eta_{ij}|D_{ij})}{\partial \eta_{ij}} = \alpha_{ij} + d_{ij} - \left(\psi_{ij} + t_{ij}\right)\exp\left\{\eta_{ij}\right\}$, and $\frac{\partial^2 \ln p(\eta_{ij}|D_{ij})}{\partial \eta_{ij}^2} = -\left(\psi_{ij} + t_{ij}\right)\exp\left\{\eta_{ij}\right\}$ respectively. From the first derivative, one can show that the mode lies at $\hat{\eta}_{ij} = \ln\left(\frac{\alpha_{ij}+d_{ij}}{\psi_{ij}+t_{ij}}\right)$; therefore the approximate normal distribution becomes

$$p\left(\eta_{ij}|D_{ij}\right) \approx N\left(\ln\left(\frac{\alpha_{ij} + d_{ij}}{\psi_{ij} + t_{ij}}\right), \frac{1}{\alpha_{ij} + d_{ij}}\right) \tag{18}$$

Once $p\left(\eta_{ij}|D_{ij}\right)$ is linearized, the linear Bayes method for linear dynamic model described in Section 2.2 of West et al. (1985) is applicable. Given the normal approximation (18), the linear Bayes method approximates the posterior distribution of θ_j by

$$q\left(\theta_j|\theta_{j-1}, D_{i,j}\right) = N\left(\mu_{ij}, C_{ij}\right),$$

where

$$\mu_{ij} = E\left[E\left[\theta_j|\,\eta_{ij}, \theta_{j-1}, D_{i,j-1}\right]|D_{i,j}\right], \tag{19}$$

and

$$C_{ij} = V\left[E\left[\theta_j|\,\eta_{ij}, \theta_{j-1}, D_{i,j-1}\right]|D_{i,j}\right] + E\left[V\left[\theta_j|\,\eta_{ij}, \theta_{j-1}, D_{i,j-1}\right]|\,D_{i,j}\right]. \tag{20}$$

The inner conditional moments in the expressions above are computed from the joint normal approximation of the prior of η_{ij} and θ_j:

$$q\left(\left(\eta_{ij},\theta_j\right)\mid\theta_{j-1},D_{i,j-1}\right)=N\left(\left(a_{ij},\theta_{j-1}\right),\begin{pmatrix}B_{ij} & A_{ij}\\ A'_{ij} & U_j\end{pmatrix}\right), \tag{21}$$

where $a_{ij}=z'_i\theta_{j-1}$, $A_{ij}=U_j\mathbf{z}_i$, $B_{ij}=\mathbf{z}'_iA_{ij}$, and U_j is the covariance matrix for the parameter evolution (7). The variance B_{ij} and covariance A_{ij} are computed from the relationship $\eta_{ij}=\mathbf{z}'_i\theta_j$. If we assume that the covariance matrix for $\theta_{j-1}|D_{j-1}$ computed from the previous interval is C_{j-1}, then according to (7) $R_j=C_{j-1}+U_j$ is the covariance matrix of $\theta_j|D_{j-1}$. Applying the discount factor method of West et al. (1985), R_j is approximated as $R_j=\frac{C_{j-1}}{\phi}$, and thus $U_j=\left(\frac{1}{\phi}-1\right)C_{j-1}$, where $0<\phi<1$ is a discount factor that controls the amount of information propagated to the subsequent interval. Note that if we use R_j in the place of U_j then we obtain the same distribution as Gamerman (1991).

The conditional expectation and variance (19)-(20) lead to

$$\begin{aligned}\mu_{ij}&=\theta_{j-1}+\frac{A_{ij}}{B_{ij}}\left(\ln\left(\frac{\alpha_{ij}+d_{ij}}{\psi_{ij}+t_{ij}}\right)-a_{ij}\right)\\ C_{ij}&=U_j-\frac{A_{ij}A'_{ij}}{B_{ij}}\left(1-\frac{1}{P_{ij}\left(\alpha_{ij}+d_{ij}\right)}\right).\end{aligned} \tag{22}$$

West et al. (1985) suggests to select the hyper-parameters α_{ij} and ψ_{ij} in order to match the true moments of the prior $p\left(\eta_{ij}|\theta_{j-1},D_{i,j-1}\right)$. This is accomplished by setting $\ln\alpha_{ij}-\ln\psi_{ij}=a_{ij}$ and $\ln\alpha_{ij}=B_{ij}$ since $P\left(\eta_{ij}|D_{i,j-1}\right)\approx N\left(\ln\left(\frac{\alpha_{ij}}{\psi_{ij}}\right),\frac{1}{\alpha_{ij}}\right)$. Hence $\alpha_{ij}=B_{ij}^{-1}$ and $\psi_{ij}=B_{ij}^{-1}\exp\left\{-a_{ij}\right\}$. Substituting these values into the expressions above yields

$$\begin{aligned}\mu_{ij}&=\theta_{j-1}+\frac{A_{ij}}{B_{ij}}\ln\frac{1+B_{ij}d_{ij}}{1+t_{ij}B_{ij}\exp\left\{a_{ij}\right\}}\\ C_{ij}&=U_j-A_{ij}A'_{ij}\left(\frac{d_{ij}}{1+d_{ij}B_{ij}}\right)\end{aligned} \tag{23}$$

Finally, the mean and covariance are updated recursively: starting from the first individual at risk μ_{1j} and C_{1j} are computed according to Equation (23), then $\mu_{ij}$and C_{ij} are updated recursively by setting $\theta_{j-1}=\mu_{i-1,,j}$ and $U_j=C_{i-1,j}$ until the last individual at risk. Finally, $\theta_j=\mu_{n_j,j}$ and $C_j=C_{n_jj}$ (for more details about this iterative procedure see Gamerman (1991); West et al. (1985)). The proposal distribution is then set as

$$q\left(\theta_j|\theta_{j-1}, D_j\right) = N\left(\mu_j,\, C_j\right) \tag{24}$$

Now, sampling from the posterior will proceed in three main steps. In the first instance, candidates are proposed from the proposal distribution (24), and in the second step weights of the proposed candidates are computed. Finally, a resampling step is performed to remove those candidates with small weights. All these steps are detailed in the following section.

3.3. Particle filter algorithm with linear Bayes tailored proposal

The following summarizes the PFLiB algorithm, which is based on the auxiliary particle filter algorithm of Pitt and Shephard (1999).

1. **Initialization:**
   Sample K particles $\left\{\theta_0^{(k)}\right\}_{k=1}^{K}$ from the prior $p\left(\theta_0\right)$ and set the weights $\left\{w_0^{(k)}\right\}_{k=1}^{K}$ to $1/K$.
2. For $j = 1, \cdots$,J
   (a) **Resampling step**
      i. Compute and normalize the sampling probabilities $\varphi_j^{(k)}$ for each particle's index $k = 1, \cdots, K$ according to Equation (13).
      ii. Sample K indexes $\left\{ind^{(k)}\right\}_{k=1}^{K}$ with probability proportional to $\varphi_j^{(k)}$, $k = 1, \cdots, K$, and remove non sampled particle paths by setting the new particle paths to $\{\theta_{0:j-1}\}_{k=1}^{K} = \left\{\theta_{0:j-1}^{ind^{(k)}}\right\}_{k=1}^{K}$, where $\theta_{0:j-1} = (\theta_0, \cdots, \theta_{j-1})$.
   (b) **Importance sampling**
      i. Based on $\theta_{j-1}^{(k)}$, compute recursively the matrix $C_j^{(k)}$, and the mean $m_j^{(k)}$ from the procedure in (23).
      ii. Sample new particles $\theta_j^{(k)} \sim q\left(\theta_j|\theta_{j-1}^{(k)}, D_j\right)$, from the importance distribution (24), and update the particle's path $\theta_{0:j}^{(k)} = \left(\theta_{0:j-1}^{(k)}, \theta_j^{(k)}\right)$, $k = 1, \cdots, K$.
      iii. Compute weights $w_j^{(k)}$ according to Equation (14).

4. Applications

We begin with a dataset on 90 patients diagnosed with gastric cancer, which was previously analysed by Gamerman (1991) and Wagner (2011). Our aim in this application is to compare the performance of the PF with the auxiliary mixture sampler (AMS) of Wagner (2011). The AMS linearizes the likelihood function (6) by means of data augmentation methods such that the dynamic survival model, defined by Equation (6) and Equation (7), reduces to a Gaussian state space model. However, the PF and the AMS are not directly comparable because the PF computes the filtering distribution while the AMS provides the smoothing distribution. In order to make the two approaches comparable, we run only the forward filtering step of the forward filtering backward sampling (FFBS) algorithm applied by Wagner (2011).

The second application consists of data on 1878 patients diagnosed with Acute Myocardial Infraction (AMI). The dataset is available in the timereg package (Scheike, 2009), a package in R under the name "TRACE". We analyse this data to assess the performance of the PFLiB in problems with several covariates.

4.1. Gastric cancer patients

This dataset consists of survival times for 90 gastric cancer patients who were assigned into two treatment groups. The first group includes survival times (in days) of 45 individuals who were treated with both chemotherapy and radiation while the second group, which includes also survival times for 45 individuals, received only a chemotherapy treatment. The only covariate included in this data is the treatment group. The data contains ten censored observations in total. The aim here is to assess the effect of combining both treatments on the risk of dying from gastric cancer.

To analyse this data, we define a dummy variable representing the treatment group ($x = 1$ for patients who received the combined treatment group, and $x = 0$ otherwise). Thus, the baseline is the chemotherapy treatment group. In order to fit the piecewise exponential model, we define the time intervals to coincide with the survival time of every third observed events. That is, each interval contains three events, except the last one. In total, there are 26 intervals. Our model has two parameter vectors in each interval: $\theta_0 = (\theta_{0,1}, \cdots, \theta_{0,26})$ representing the log baseline hazard, and

$\theta_1 = (\theta_{1,1}, \cdots, \theta_{1,26})$ representing the combined treatments effect. Alternative time partitions are possible and once could in principle compare different partitions by comparing their marginal likelihoods.

In order to compare our method with the AMS sampler, we set, for both methods, the initial distribution $p(\theta_{01}, \theta_{11})$ to the bivariate normal distribution with mean zero, and a diagonal covariance matrix where the variance of each covariate is set to 100. For the AMS, we perform 50000 iterations with a burn-in period of 25000. Following Wagner (2011) all hyperparameter values for the inverse gamma prior process defined for the prior (7) are set to 0.01. In order to have the same number of posterior simulations for both methods, we set the number of particles in the PFLiB algorithm to 25000. The performance of the PFLiB depends largely on the discount factor ϕ. We select the discount factor based on the Watanabe Akaike information criteria (WAIC), following Equations 12 and 13 in Gelman et al. (2014). The lowest WAIC (1162) is obtained when $\phi = 0.4$.

Figure 1 presents the resulting posterior mean trajectories of the effect parameter θ_1 and the log baseline hazard θ_0 together with their corresponding 95% highest probability density (HPD) intervals obtained from the PFLiB and AMS algorithms.

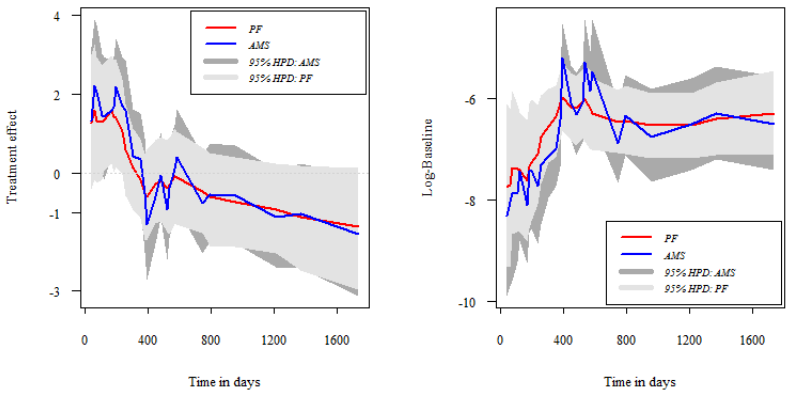

Figure 1: Gastric cancer data: Mean trajectory for the effect parameter of the combined treatments (left), and the log baseline hazard (right) together with their corresponding 95% HPD.

From Figure 1 we observe that the effect of the combined treatments on

the risk of death is positive but decreasing during the first 400 days. After this period, the combined treatments becomes negative and continues to decrease steadily. On the other hand the log baseline hazard increases within the first 400 days of the treatment, and then stabilizes afterwards. However, the trajectory obtained from the auxiliary sampler shows a sharper decline in the log baseline hazard before it stabilizes.

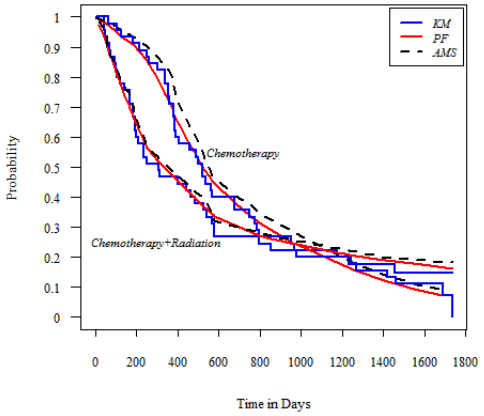

Figure 2: Gastric cancer data: Kaplan Meier estimates (blue step function), and survival curves estimated by the PFLiB (red line), and the AMS (black broken line).

Figure 2 displays the Kaplan Meier survival curves together with the survival curves estimated by PFLiB and AMS methods. We observe that, at least for the first 1000 days, patients treated with chemotherapy (baseline) have a higher survival compared to those treated with both radiography and chemotherapy. Combining treatments show effect after 1000 days. The median survival time of patients in the combined treatments group is about 400 days while for the chemoterapy treatment group it's about 600 days. We also note that the Kaplan Meier estimates are slightly more closer to the survival curves (and especially the baseline curves) estimated from the Particle Filter algorithm than to those estimated by the AMS.

To compare the computation time of the PFLiB algorithm to that of the AMS, we use a built-in function in R (system.time) on a windows computer with 8 RAM memory, and an i7-4790 3.60 GHz processor. We observe that the PFLiB is about 11 times faster than the AMS; 50000 iterations took

135 minutes (2.25 hours) for the AMS, and 12.28 minutes for the PFLiB. This is not surprising because the AMS involves many computations. For instance, the linearization step requires sampling auxiliary survival times for censored observations, sampling latent component allocation indicators from the mixture of ten Gaussian components for all observations in the risk set, and then computing the Kalman filter moments.

Further, we compared the convergence of both methods through the average number of effective sample size per minute. We average the effective sample size per minute for all parameters in the model: The average effective sample size per minute for the AMS is 2.052 and for the PFLiB is 586.669. This is an indication that the PFLiB explore the posterior distribution more efficiently than to the AMS.

4.2. Patients with Acute Myocardial Infraction (AMI).

In this illustration we apply the PFLiB algorithm to assess the effects of more than one risk factors (covariates) on the survival time of patients who were diagnosed with AMI. The data contains survival times of 1878 patients with Acute Myocardial Infraction (AMI). Furthermore, it contains different risk factors- patient's heart pumping measured in ultrasound measurements (wmi), patient's ventricular fibrillation (vf) where $vf = 1$: presence and $vf = 0$ indicates absence, patient's clinical heart pump failure (chf) where $chf = 1$ indicates presence and $chf = 0$ indicates absence, indicator of patient's diabetes (1: present, 0: absent), and patient's sex (1: female, 0: male). The aim of the study is to estimate the effect of various risk factors on the patients' survival time.

By the end of the follow-up time, 970 of the 1878 patients (52%) have died from myocardial infraction while the rest 908 (48%) were still alive or died from other causes and, hence, were considered as censored. The survival time (in years) was partitioned such that each interval contained 40 events, which resulted in 24 intervals ($970/40 \approx 24$). Thus, the model includes 24 sets of parameters corresponding to the number of intervals, and each parameter set is a vector of length six corresponding to the log-baseline (intercept) and five risk factors.

To run the PFLiB algorithm, we set the initial distribution $p(\theta_0)$ to the multivariate normal distribution with mean zero, and a diagonal covariance matrix where the variance of each covariate is set to 100. Further, we set the number of particles to 30000. We fitted the model with the discount factor $\phi = 0.4$, because the lowest WAIC were obtained at this value of ϕ.

The posterior estimates for the parameters in the model are presented in Figure 3. One can see that the log baseline hazard (intercept), as well as the risk factors *wmi*, *chf* and the *vf* have dynamic trends. The log baseline hazard decreases sharply in the first year and remains nearly constant for the next three years which is followed by a slight linear increase. The effect of the *wmi* increases in the first four years and then stabilizes. However, it is always negative indicating that the hazards of dying from AMI decreases with increasing *wmi* values. The effect of *chf* is around one (implying patients with *chf* die from AMI at a rate of exp(1)= 2.7 times those without *chf*) during the first five years after which it drops to 0.4 (corresponding to relative hazard of 1.49). The effect of *vf* is positive in the first year and stabilizes around -0.6 afterwards. One can also note that the effects of diabetes and sex are nearly constant since their mean trajectories are more or less horizontal throughout the study period. Generally, females seem to have lower rates of death from AMI than males. Lastly, patients with diabetes seem to have an elevated risk of dying from AMI compared to diabets-free patients.

In accordance with Jensen et al. (1997), who previously analysed the data, our results indicate that *chf* and *vf* are the most significant risk factors of death from AMI. For this reason, we plot the survival curves for the groups corresponding to combinations of these two factors. The survival curves estimated from the PFLiB (smooth curves) and empirical survival function estimated using Kaplan Meier (step functions) are presented in Figure 4. The group of patients with $chf = 0$ and $vf = 0$ have the highest survival curve, followed by the group with $chf = 1$ and $vf = 0$. The lowest survival curve corresponds to the group with $chf = 1$ and $vf = 1$. Thus, patient with both ventricular fibrillation (vf) and clinical heart pump failure (chf) have the highest risk of death from AMI.

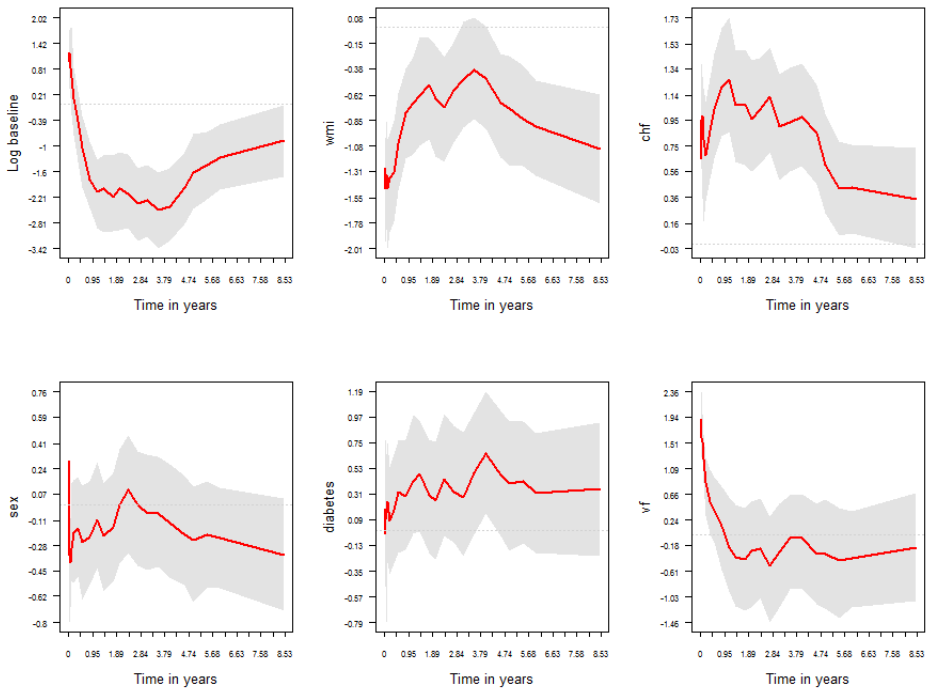

Figure 3: AMI data: Posterior mean trajectories of the effect parameters and the log baseline hazards (intercept) with their corresponding 95% HPD (grey band)

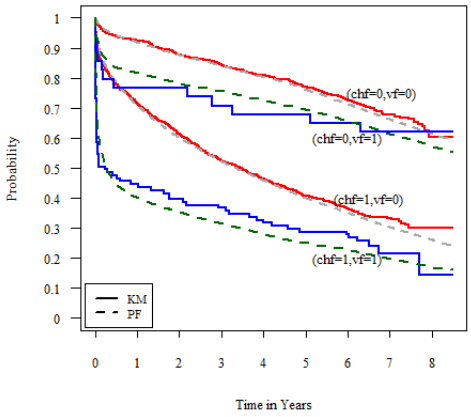

Figure 4: AMI data: Survival curves of the groups corresponding to the combination of different levels of *chf* and *vf*

5. Concluding remarks

Dynamic survival models extend the standard proportional hazard models in order to allow for covariate-effects to vary over time. Such models are naturally handled with a Bayesian approach which allows flexibility in incorporating past information and advanced relationships among parameters through the prior distribution. However, the models are generally complex due to nonlinearity and high correlations among the parameters in the model.

In the present work, we apply Sequential Monte Carlo methods known as the Particle Filter to handle the above problems. This method requires designing a good proposal distribution; we developed it by approximating the posterior distribution with a second order Taylor series expansion and applying the linear Bayes method. We illustrated the method with two datasets containing one and five covariates, respectively.

Our approach replicates previous results obtained through MCMC sampling methods, but in a much more efficient way. A comparison of our approach with the auxiliary mixture sampler shows that the particle filter is much faster and explores the posterior more efficiently as indicated by a significantly higher average effective sample size per minute. Further, the particle filter performs very well on a dataset with more than one covariate.

A major drawback of our approach is that the particle filter can only compute the filtering distribution. As smoothing distribution is needed in order to make predictions, we suggest to apply a particle smoothing algorithm (Briers et al., 2010) that enables sampling from the smoothed posterior distribution. Another possible future extension is to incorporate in the model time-varying covariates (Houwelingen, 2007) or anticipatory covariates (Ghilagaber and Koskinen, 2009).

Acknowledgements: I would like to thank Ludwig Fahrmeir and Mattias Villani for their constructive comments, and Helga Wagner for sharing the code for the auxiliary mixture sampler.

References

Andrieu, C., Doucet, A., Holenstein, R., 2010. Particle markov chain monte carlo methods. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72 (3), 269–342.

Arulampalam, M. S., Maskell, S., Gordon, N., Clapp, T., 2002. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE Transactions on signal processing 50 (2), 174–188.

Briers, M., Doucet, A., Maskell, S., 2010. Smoothing algorithms for state–space models. Annals of the Institute of Statistical Mathematics 62 (1), 61–89.

Chopin, N., Jacob, P. E., Papaspiliopoulos, O., 2013. Smc2: an efficient algorithm for sequential analysis of state space models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75 (3), 397–426.

Cox, R. D., 1972. Regression models and life tables (with discussion). Journal of the Royal Statistical Society 34, 187–220.

Das, S., Dey, D. K., 2013. On dynamic generalized linear models with applications. Methodology and Computing in Applied Probability, 1–15.

Del Moral, P., Doucet, A., Jasra, A., 2006. Sequential monte carlo samplers. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68 (3), 411–436.

Doucet, A., De Freitas, N., Gordon, N., 2001. An introduction to sequential Monte Carlo methods. In: Sequential Monte Carlo methods in practice. Springer, pp. 3–14.

Doucet, A., Godsill, S., Andrieu, C., 2000. On sequential Monte Carlo sampling methods for Bayesian filtering. Statistics and computing 10 (3), 197–208.

Fahrmeir, L., 1994. Dynamic modelling and penalized likelihood estimation for discrete time survival data. Biometrika 81 (2), 317–330.

Fahrmeir, L., Kneib, T., 2011. Bayesian smoothing and regression for longitudinal, spatial and event history data. Oxford University Press, New York.

Fahrmeir, L., Wagenpfeil, S., 1996. Smoothing hazard functions and time-varying effects in discrete duration and competing risks models. Journal of the American Statistical Association 91 (436), 1584–1594.

Fearnhead, P., Wyncoll, D., Tawn, J., 2010. A sequential smoothing algorithm with linear computational cost. Biometrika 97 (2), 447–464.

Frühwirth-Schnatter, S., 1994. Data augmentation and dynamic linear models. Journal of time series analysis 15 (2), 183–202.

Gamerman, D., 1991. Dynamic Bayesian models for survival data. Applied Statistics, 63–79.

Gamerman, D., 1998. Markov Chain Monte Carlo for dynamic generalised linear models. Biometrika 85 (1), 215–227.

Gelman, A., Hwang, J., Vehtari, A., 2014. Understanding predictive information criteria for Bayesian models. Statistics and Computing 24 (6), 997–1016.

Ghilagaber, G., Koskinen, J., 2009. Bayesian Adjustment of Anticipatory Covariates in Analyzing Retrospective Data. Mathematical Population Studies 16 (2), 105–130.

Hemming, K., Shaw, J., 2002. A parametric dynamic survival model applied to breast cancer survival times. Journal of the Royal Statistical Society: Series C (Applied Statistics) 51 (4), 421–435.

Hemming, K., Shaw, J., 2005. A class of parametric dynamic survival models. Lifetime data analysis 11 (1), 81–98.

Hennerfeind, A., Brezger, A., Fahrmeir, L., 2006. Geoadditive survival models. Journal of the American Statistical Association 101 (475), 1065–1075.

Houwelingen, H. C. V., 2007. Dynamic prediction by landmarking in event history analysis. Scandinavian Journal of Statistics 34 (1), 70–85.

Houwelingen, H. C. V., Putter, H., 2011. Dynamic prediction in clinical survival analysis. CRC Press.

Jensen, G., Torp-Pedersen, C., Hildebrandt, P., Kober, L., Nielsen, F., Melchior, T., Joen, T., Andersen, P., 1997. Does in-hospital ventricular fibrillation affect prognosis after myocardial infarction? European Heart Journal 18 (6), 919–924.

Kalman, R. E., 1960. A new approach to linear filtering and prediction problems. Journal of Basic Engineering 82 (1), 35–45.

Lindsten, F., Jordan, M. I., Schön, T. B., 2014. Particle gibbs with ancestor sampling. Journal of Machine Learning Research 15 (1), 2145–2184.

Liu, J. S., Chen, R., 1998. Sequential Monte Carlo methods for dynamic systems. Journal of the American Statistical Association 93 (443), 1032–1044.

Migon, H. S., Gamerman, D., Lopes, H. F., Ferreira, M. A., 2005. Dynamic models. Handbook of Statistics 25, 553–588.

Pitt, M. K., Shephard, N., 1999. Filtering via simulation: Auxiliary particle filters. Journal of the American Statistical Association 94 (446), 590–599.

Sargent, D. J., 1997. A flexible approach to time-varying coefficients in the Cox regression setting. Lifetime data analysis 3 (1), 13–25.

Scheike, T., 2009. Timereg: timereg package for flexible regression models for survival data. R package version 1.2-4.

Van Der Merwe, R., Doucet, A., De Freitas, N., Wan, E. A., 2001. The unscented particle filter. In: Advances in neural information processing systems. pp. 584–590.

Wagner, H., 2011. Bayesian estimation and stochastic model specification search for dynamic survival models. Statistics and Computing 21 (2), 231–246.

West, M., Harrison, P. J., Migon, H. S., 1985. Dynamic generalized linear models and Bayesian forecasting. Journal of the American Statistical Association 80 (389), 73–83.