

Latent diffusion models for survival analysis

Gareth O. Roberts^a and Laura M. Sangalli^{b*}

^a*CRiSM, Department of Statistics, University of Warwick, Coventry CV4 7AL, UK*

^{b*}*MOX, Dipartimento di Matematica, Politecnico di Milano, P.zza L. da Vinci 32, 20133 Milano, Italy*

Abstract

We consider Bayesian hierarchical models for survival analysis, where the survival times are modeled through an underlying diffusion process, which determines the hazard rate. We show how these models can be efficiently treated by means of Markov chain Monte Carlo techniques.

Keywords and phrases. Diffusion processes, survival analysis, parametrization of hierarchical models.

1 Introduction

Diffusion processes have found many applications in the modelling of continuous-time phenomena, for problems related to several scientific areas, ranging from economics to biology, from physics to engineering. Here we use diffusion processes as building blocks for the definition of models for survival and event history analysis. This idea is not new (see for example the reviews in Aalen and Gjessing (2001, 2004)). However, in this paper we are able to considerably extend the flexibility of the diffusion models used, by adopting powerful Markov Chain Monte Carlo techniques.

Diffusion models for survival analysis have been proposed because, as summarized in Aalen and Gjessing (2004), “when modelling survival data it may be of interest to imagine an underlying process leading up to the event in question”. Such a process might for example represent the development of a disease. Two types of models have been considered in the literature. Models where the event happens when a diffusion process hits some barrier, and models where the hazard rate is some suitable function of the diffusion. For the former type of models, we refer the reader to Aalen and Gjessing (2001), Aalen, Borgan, and Gjessing (2008), and references therein. Here we are interested in the latter. Woodbury and Manton (1977) proposed a model where the hazard rate is a quadratic function of an Ornstein-Uhlenbeck diffusion process. This model has been later considered by several authors, including Myers (1981), Yashin (1985), Yashin and Vaupel (1986), and Aalen and Gjessing (2004). For given values of the parameters of the Ornstein-Uhlenbeck process, survival distributions and hazards are studied. Myers (1981) focuses on survival distributions conditioned on initial covariates values; Yashin (1985) and Yashin and Vaupel (1986) use hazards based on quadratic functions of Ornstein-Uhlenbeck processes in order to model heterogeneity among groups and among individuals, and study the relative hazard functions and survival distributions; Aalen and Gjessing (2004) derive quasi-stationary distributions. Obtaining such analytical results for hazard functions other than quadratic functions, or for more complex diffusion

processes, is not feasible.

In our paper, we adopt a Bayesian approach and we show how these models can be efficiently treated by means of Markov chain Monte Carlo techniques, for general choices of diffusion processes and hazard functions. For instance, by the proposed methods it is possible to deal with latent diffusion models which are stochastic perturbations of common survival models. We also consider the case of multiple groups of observations, typical of clinical trials, and we show how to efficiently deal with covariates. We illustrate the methods via simulation studies and applications to real data.

It should be mentioned that other classes of Bayesian nonparametric and semiparametric models for survival analysis have been proposed in literature. Among the most important, we mention the models based on *neutral to the right random probabilities*, whose cumulative hazard rates are processes with independent increments (see Doksum (1974) and Ferguson (1974) for definition and properties of these random measures, and e.g. Susarla and Van Ryzin (1976), Kalbfleisch (1978), Ferguson and Phadia (1979), Hjort (1990) and Damien and Walker (2002) for applications in survival analysis), and all models falling within the framework of *multiplicative intensity models*, whose hazard rates are mixtures of known kernels where the mixing measure is a weighted gamma process (see Dykstra and Laud (1981), Lo and Weng (1989), Ishwaran and James (2004) and references therein).

The paper is organized as follows. In Section 2 we recall the essential of diffusion processes and introduce the model; we also outline how, in the described framework, it is possible to consider stochastic perturbations of common survival models. In Section 3 we describe the MCMC scheme and gives the details of a suitable Hastings-within-Gibbs algorithm, showing its implementation by means of a toy example. In Section 4 we present improved versions of the algorithm, based on reparametrizations of the model. In Section 5 we discuss a straightforward generalization of the framework developed in the previous sections, and deal with the case of multiple groups of observations; this is also illustrated by application to a dataset from a clinical trial, that has been considered in a number of papers in the context of survival analysis, the famous Cox (1972) paper among the firsts. In Section 6 we describe how covariates can be efficiently included in the proposed models, and give an illustrative application to the lung-cancer dataset analysed by Muers et al. (1996). Finally, in Sections 7 and 8 we discuss possible extensions of the models considered.

2 Latent diffusion models

Let Θ be a random variable with values in \mathbb{R}^d . Denote by $C([0, \infty), \mathbb{R})$ the space of continuous functions from $[0, \infty)$ to \mathbb{R} , and by \mathcal{C} its cylinder σ -algebra. Given $\Theta = \theta$, consider the scalar diffusion process $X = \{X_t : t \geq 0\}$, solution of a *stochastic differential equation* (SDE, for short) of the form

$$\begin{aligned} dX_t &= \beta(X_t, \theta) dt + \sigma dB_t & t \geq 0 \\ X_0 &= x_0 \end{aligned} \tag{1}$$

driven by the standard scalar Brownian motion $B = \{B_t : t \geq 0\}$. The Brownian motion B and the diffusion process X are random elements of $(C([0, \infty), \mathbb{R}), \mathcal{C})$. The diffusion coefficient σ is assumed constant and known, for the moment. The more technically difficult case of unknown σ

is postponed to Section 7. The drift $\beta(x, \theta)$ is assumed to be jointly measurable in x and θ , and to satisfy the regularity conditions (locally Lipschitz, with linear growth bound) that guarantee the existence of a weakly unique global solution to (1). See, for example, Chapter V.24 in Rogers and Williams (2000).

Let \mathbb{W}_σ be the law of σB , and, for a given θ , denote by \mathbb{P}_θ the law of the diffusion X , solution of (1). By *Girsanov's theorem*, the Radon-Nikodym derivative of \mathbb{P}_θ , with respect to \mathbb{W}_σ , is given by

$$\frac{d\mathbb{P}_\theta}{d\mathbb{W}_\sigma}(x) = \exp \left\{ \int_0^\infty \frac{\beta(x_t, \theta)}{\sigma^2} dx_t - \frac{1}{2} \int_0^\infty \frac{\beta(x_t, \theta)^2}{\sigma^2} dt \right\}$$

where x is an element of $(C([0, \infty), \mathbb{R}), \mathcal{C})$. See, for example, Chapter V.27 in Rogers and Williams (2000).

Similarly, for a finite T , denote by $C([0, T], \mathbb{R})$ the space of continuous functions from $[0, T]$ to \mathbb{R} , and by \mathcal{C}^T its cylinder σ -algebra. Then, $B_{[0, T]} := \{B_t : 0 \leq t \leq T\}$ and $X_{[0, T]} = \{X_t : 0 \leq t \leq T\}$ are random elements of $(C([0, T], \mathbb{R}), \mathcal{C}^T)$. Let $\mathbb{W}_{T, \sigma}$ be the law of $\sigma B_{[0, T]}$, and, for a given θ , denote by $\mathbb{P}_{T, \theta}$ the law of $X_{[0, T]}$. Then, by Girsanov's theorem, the Radon-Nikodym derivative of $\mathbb{P}_{T, \theta}$, with respect to $\mathbb{W}_{T, \sigma}$, is given by

$$\frac{d\mathbb{P}_{T, \theta}}{d\mathbb{W}_{T, \sigma}}(x_{[0, T]}) = \exp \left\{ \int_0^T \frac{\beta(x_t, \theta)}{\sigma^2} dx_t - \frac{1}{2} \int_0^T \frac{\beta(x_t, \theta)^2}{\sigma^2} dt \right\} \quad (2)$$

and, for each T , the measures $\mathbb{P}_{T, \theta}$ are absolutely continuous.

Given the diffusion X , let us consider the random distribution function $F_{X, h}$ on $[0, \infty)$, defined as

$$F_{X, h}(t) := 1 - \exp \left\{ - \int_0^t h(X_s) ds \right\} \quad t \geq 0 \quad (3)$$

where $h(\cdot)$ is some suitable nonnegative and continuous function, with $\int_0^\infty h(X_s) ds = \infty$ almost surely. The function $h(\cdot)$ plays the role of the hazard function, and $h(X_t)$ is the random hazard rate, at time t , associated to the random distribution $F_{X, h}$.

Two features of the random measure $F_{X, h}$ have to be noted. The first is that the hazard inherits the Markov property of the diffusion process, so that the hazard at a future time t' just depends on the hazard at the present time t . The Markov property seems indeed a sensible choice to make at the level of the hazard. The second is that the cumulative hazard is a process with positively correlated increments, being the integral of a continuous process. The latter feature is natural in many contexts, and it inserts in the model the concern with the stochastic process that clearly must lie behind the occurrence of events. In words, an high increment of the cumulative hazard over the time interval $[t, t']$ means that the underlying stochastic process has reached a region of high risk, and this is likely to yield an high increment of the cumulative hazard over a close (disjoint) time interval. The strength of this positive correlation, and thus the smoothness of the cumulative hazard, depends on the choice of the hazard function h and of the diffusion process X : the rougher the diffusion, the weaker is the correlation, and viceversa. See also the comments in Section 8. Note that the property we have just highlighted differentiates the models we are considering from models based on neutral to the right random probabilities, whose cumulative hazards are processes with independent increments and thus have an erratic behaviour.

Let us now consider a sequence of event times Y_1, Y_2, \dots which are, conditionally on $F_{X,h}$, independent and identically distributed (i.i.d., for short) with common distribution $F_{X,h}$. From (3), it follows that the distribution of Y_1, \dots, Y_n , given $X = x$, has density, with respect to the n -dimensional Lebesgue measure \mathcal{L}^n , given by

$$l(y_1, \dots, y_n | x) := \left[\prod_{j=1}^n h(x_{y_j}) \right] \exp \left\{ - \sum_{j=1}^n \int_0^{y_j} h(x_t) dt \right\}. \quad (4)$$

Censored observations can be easily dealt with in this setting. In the present paper, we shall restrict our attention to independent right-censored schemes. Let (y_1, \dots, y_m) be the observed event times and let $(y_{m+1+}, \dots, y_{n+})$ be the right-censored event times, then the likelihood becomes

$$\begin{aligned} & l(y_1, \dots, y_m, y_{m+1+}, \dots, y_{n+} | x) \\ &= \left[\prod_{j=1}^m h(x_{y_j}) \right] \exp \left\{ - \sum_{j=1}^m \int_0^{y_j} h(x_t) dt - \sum_{j=m+1}^n \int_0^{y_{j+}} h(x_t) dt \right\}. \end{aligned}$$

We are thus considering a latent diffusion model for survival analysis, where the survival times are modelled through an underlying diffusion process which determines the hazard rate. As highlighted by Aalen and Gjessing (2004), this model can be also interpreted as a random barrier hitting model. Indeed, the event happens when the cumulative hazard strike a random barrier R , which is exponentially distributed with mean 1, and is stochastically independent of X .

2.1 Stochastic perturbations of common survival models

In the framework we have described, one possibility is to consider stochastic perturbations of common survival models. Heuristically, the idea is that if we can express the hazard $r(t)$ of a given model as a solution of an ordinal differential equation $\frac{dr(t)}{dt} = g(r(t))$, for some suitable function g , then we may be able to use g , or some modification of it, to model the drift of a SDE; starting from this SDE we can thus consider a latent diffusion model whose hazard function is a stochastic perturbation of $r(t)$.

We shall illustrate this by some examples. The simplest case is offered by the Gompertz model. The Gompertz hazard $r(t) = \beta \exp\{\alpha t\}$, for $\alpha, \beta > 0$, is a solution of the ordinal differential equation $\frac{dr(t)}{dt} = g(r(t)) = \alpha r(t)$. Consider thus the latent diffusion model based on the SDE having drift $g(X_t) = \theta X_t$, for $\theta > 0$,

$$dX_t = \theta X_t dt + \sigma dB_t, \quad t \geq 0, \quad X_0 = x_0 > 0, \quad (5)$$

and with hazard function $h(u) = |u|$. For $\sigma = 0$, the SDE (5) reduces to the ordinal differential equation written above, for which the Gompertz hazard is a solution, and the latent diffusion model reduce to the Gompertz model. Hence, the latent diffusion model based on the SDE (5), with hazard function $h(u) = |u|$, can be seen as a stochastic perturbation around a central Gompertz model. This constitutes a simple example of latent diffusion model, for which the law of X_t is known, and thus also the law of the hazard. In the other examples we shall now give, the SDE can not be explicitly solved, but the latent diffusion models based on them can be treated by the techniques described in the present paper.

Let us consider the Weibull model, whose hazard $r(t) = \alpha \beta t^{\alpha-1}$, for $\alpha, \beta > 0$, is a non-trivial solution of the ordinal differential equation $\frac{dr(t)}{dt} = g(r(t)) = \gamma r(t)^{(\alpha-2)/(\alpha-1)}$. Consider thus the

latent diffusion model based on the SDE

$$dX_t = \theta_1 (\text{sign}(X_t)) |X_t|^{\theta_2} dt + \sigma dB_t, \quad t \geq 0, \quad X_0 = x_0 > 0, \quad (6)$$

where

$$\text{sign}(u) = \begin{cases} 1 & \text{if } u > 0 \\ -1 & \text{if } u < 0 \\ 0 & \text{if } u = 0 \end{cases}$$

and with hazard function $h(u) = |u|$. For $\sigma = 0$, the SDE (6) reduces to the ordinal differential equation written above, for which the Weibull hazard is a solution (θ_2 plays here the role of $(\alpha - 2)/(\alpha - 1)$). Hence, the latent diffusion model based on the SDE (6), with hazard function $h(u) = |u|$, can be seen as a stochastic perturbation around a central Weibull model. For values of θ_2 in the interval $(0, 1)$, which correspond to $\alpha > 2$, the SDE (6) has a non-explosive solution. This solution is weakly unique (see e.g. Stroock and Varadhan (2006)). In Sections 5.1 and 6.1 we shall implement this latent diffusion model in some illustrative applications to real data.

Using the simple idea outlined above, it is possible to develop other latent diffusion models, such as stochastic perturbations of Log-logistic models and Exponential-power models. The Log-logistic hazard ($r(t) = \alpha\beta t^{\alpha-1}/(1 + \beta t^\alpha)$, for $\alpha, \beta > 0$) and the Exponential-power hazard ($r(t) = \alpha\beta^\alpha t^{\alpha-1} \exp\{-(\beta t)^\alpha\}$, for $\alpha, \beta > 0$) can in fact be written as solutions of $\frac{dr(t)}{dt} = g(r(t))$, for suitable functions g (when $\alpha < 1$ for the Log-logistic, and $\alpha > 1$ for the Exponential-power). Let us give a further example, which generalises the Pareto model. The Pareto hazard $r(t) = \alpha/t$, for $\alpha > 0$ and $t \geq \lambda > 0$, is a solution of the equation $\frac{dr(t)}{dt} = g(r(t)) = -\frac{1}{\alpha}[r(t)]^2$. Now, the SDE having drift $g(X_t) = -\theta X_t^2$, for $\theta > 0$,

$$dX_t = -\theta X_t^2 dt + \sigma dB_t, \quad t \geq \lambda > 0, \quad X_\lambda = x_\lambda > 0, \quad (7)$$

provides a stochastic perturbation around the Pareto hazard, but unfortunately, this SDE cannot be used for our purposes since it has an explosive solution. On the other hand, we can modify (7), for example by inclusion of X_t in the diffusion coefficient, in order to obtain another SDE

$$dX_t = -\theta X_t^2 dt + \sigma X_t dB_t, \quad t \geq \lambda > 0, \quad X_\lambda = x_\lambda > 0, \quad (8)$$

that also provides a stochastic perturbation around the Pareto hazard, but has a non-explosive solution. The latter SDE can thus be transformed into one of constant diffusion coefficient, which can in turn be used in the latent diffusion model; note that the solution of (8), and of the corresponding SDE with constant coefficient, are almost surely positive, and thus we can take as hazard function $h(\cdot)$ the identity function, obtaining a particularly natural perturbation of the Pareto. It is worth recalling that an SDE with general diffusion coefficient $\sigma(X_t, \theta)$,

$$dX_t = \beta(X_t, \theta) dt + \sigma(X_t, \theta) dB_t, \quad t \geq 0, \quad X_0 = x_0,$$

can in fact be transformed into an SDE of unit diffusion coefficient for the process Y , by applying the 1-1 transformation $X_t \rightarrow \eta(X_t; \theta) =: Y_t$, where $\eta(x; \theta) = \int^x \frac{1}{\sigma(z; \theta)} dz$ is any anti-derivative of $\sigma^{-1}(\cdot; \theta)$ (we are assuming $\sigma(x, \theta)$ is differentiable for any $x \in C([0, \infty), \mathbb{R})$). See e.g. Beskos et al. (2006). This approach opens up to a number of possible stochastic perturbations of commonly used hazards.

3 Markov Chain Monte Carlo methods for latent diffusion models

Let $p_\Theta(\theta)$ be the prior density, with respect to \mathcal{L}^d , of the d -dimensional parameter Θ , which appears in the drift of the diffusion process X , solution of (1). Fix a finite time horizon T of interest, with $T \geq y_{[n]}$, where $y_{[n]} := \max\{y_1, \dots, y_n\}$. The choice of T will be discussed in Section 4. Then, the joint posterior distribution of Θ and $X_{[0,T]}$ has density, with respect to the product measure $\mathcal{L}^d \otimes \mathbb{W}_{T,\sigma}$, given by

$$\pi(\theta, x_{[0,T]} | y_1, \dots, y_n) = C p_\Theta(\theta) g(x_{[0,T]} | \theta) l(y_1, \dots, y_n | x_{[0,y_{[n]}]}) \quad (9)$$

where C is a normalizing constant, and $g(x_{[0,T]} | \theta) := \frac{d\mathbb{P}_{T,\theta}}{d\mathbb{W}_{T,\sigma}}(x)$ is given by Girsanov's formula (2).

A Gibbs sampling algorithm for sampling from (9) alternates between

1. simulation of Θ , conditional on the observations and the current path of $X_{[0,T]}$;
2. simulation of $X_{[0,T]}$, conditional on the observations and the current value of Θ .

Note that the parameter Θ and the observations Y_1, \dots, Y_n are conditionally independent, given the non-observed process $X_{[0,T]}$. In particular, from (9), the conditional distribution of Θ given $X_{[0,T]}$, has density, with respect to \mathcal{L}^d , proportional to $p_\Theta(\theta) g(x_{[0,T]} | \theta)$. The update of the parameter is particularly straightforward when a conjugate prior $p_\Theta(\theta)$ is chosen, so that it is possible to derive analytically the conditional distribution of Θ given $X_{[0,T]}$ and sample directly from it. The second step is computationally more demanding. From (9), the conditional distribution of $X_{[0,T]}$, given parameter and observations, has density, with respect to $\mathbb{W}_{T,\sigma}$, proportional to $g(x_{[0,T]} | \theta) l(y_1, \dots, y_n | x)$, and cannot be sampled directly. An appropriate Metropolis-Hastings step is thus required.

Implementation of the algorithm will necessarily involve a discretisation of the diffusion sample path. When the SDE cannot be solved, it is possible to use *Euler-Maruyama approximation*. See for example Chapter 9 in Kloeden and Platen (1992). Alternatively, it may be possible to simulate the diffusion path by means of the exact algorithm described in Beskos, Papaspiliopoulos, Roberts, and Fearnhead (2006), thus avoiding approximation errors.

3.1 Hastings-within-Gibbs algorithm for a latent diffusion model

We now give the details of the Hastings-within-Gibbs algorithm for latent diffusion models.

Just as an example, consider a latent diffusion model with base diffusion which is solution of the SDE

$$dX_t = \theta^\top f(X_t) dt + \sigma dB_t, \quad t \geq 0, \quad X_0 = x_0, \quad (10)$$

with $\theta^\top = (\theta_1, \dots, \theta_d)$, and $f(x)^\top = (f_1(x), \dots, f_d(x))$, where $f_i(x)$ is some real-valued function, for $i = 1, \dots, d$. Let the drift $\theta^\top f(x)$ be such that the regularity conditions mentioned in Section 2 are satisfied. Let the prior for $\Theta = (\Theta_1, \dots, \Theta_d)$ be multivariate Gaussian, with mean vector and

variance matrix

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_d \end{bmatrix} \quad \Sigma = \begin{bmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1d} \\ \lambda_{12} & \lambda_{22} & \cdots & \lambda_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{1d} & \lambda_{2d} & \cdots & \lambda_{dd} \end{bmatrix}^{-1}$$

Then, the distribution of Θ , given the diffusion $X_{[0,T]} = x_{[0,T]}$, is still Gaussian, with mean and covariance matrix

$$\mu_x = \Sigma_x \begin{bmatrix} S_1 \\ S_2 \\ \vdots \\ S_d \end{bmatrix} \quad \Sigma_x = \begin{bmatrix} L_{11} & L_{12} & \cdots & L_{1d} \\ L_{12} & L_{22} & \cdots & L_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ L_{1d} & L_{2d} & \cdots & L_{dd} \end{bmatrix}^{-1} \quad (11)$$

where, for $i = 1, \dots, d$ and $j = 1, \dots, d$,

$$S_i := \frac{1}{\sigma^2} \int_0^T f_i(x_t) dx_t + \sum_{j=1}^d \lambda_{ij} \mu_j \quad L_{ij} := \frac{1}{\sigma^2} \int_0^T f_i(x_t) f_j(x_t) dt + \lambda_{ij}.$$

The update of Θ can thus be performed by sampling directly from this conditional distribution.

The update of the diffusion $X_{[0,T]}$ is less straightforward and requires an appropriate Metropolis-Hastings step. It is possible for example to carry out an independence sampler with proposal distribution given by a Brownian motion starting at x_0 . To improve the acceptance rate of the move that update the diffusion path, we apply the following updating strategy. Let $0 = t_1 < \dots < t_m = T$. Instead of proposing a new diffusion path on the whole interval $[0, T]$, we propose to change the trajectory just on a subinterval $[t_i, t_{i+2}]$, keeping fixed the rest of the diffusion. To ensure continuity of the diffusion path, the proposal distribution, for the new trajectory on the subinterval $[t_i, t_{i+2}]$, is a Brownian bridge $BB_{[t_i, t_{i+2}]}(x_{t_i}, x_{t_{i+2}}) = \{BB_t(x_{t_i}, x_{t_{i+2}}) : t_i \leq t \leq t_{i+2}\}$, having as starting and ending points, respectively, the values $X_{t_i} = x_{t_i}$ and $X_{t_{i+2}} = x_{t_{i+2}}$ of the current diffusion. The proposed diffusion path $x_{[0,T]}^*$ is then given by $\{x_t^* = 1(t \notin [t_i, t_{i+2}])x_t + 1(t \in [t_i, t_{i+2}])bb_t(x_{t_i}, x_{t_{i+2}}) : t \in [0, T]\}$, where $bb_t(x_{t_i}, x_{t_{i+2}})$ is the realization of the Brownian bridge $BB_{[t_i, t_{i+2}]}(x_{t_i}, x_{t_{i+2}})$. This move is accepted with probability

$$1 \wedge \frac{g(bb_{[t_i, t_{i+2}]}(x_{t_i}, x_{t_{i+2}})|\theta)}{g(x_{[t_i, t_{i+2}]}|\theta)} \frac{l(y_1, \dots, y_n | x_{[0, y_{[n]]}^*})}{l(y_1, \dots, y_n | x_{[0, y_{[n]]})} \quad (12)$$

where $g(x_{[t_i, t_{i+2}]}|\theta)$ is given by Girsanov's formula restricted to the interval $[t_i, t_{i+2}]$, i.e.

$$g(x_{[t_i, t_{i+2}]}|\theta) = \exp \left\{ \int_{t_i}^{t_{i+2}} \frac{\theta^\top f(X_t)}{\sigma^2} dx_t - \frac{1}{2} \int_{t_i}^{t_{i+2}} \frac{(\theta^\top f(X_t))^2}{\sigma^2} dt \right\}.$$

The procedure is iterated for $i = 1, \dots, m-3$. Note that the different blocks $[t_i, t_{i+2}]$ overlap, so that there are no time instants where the diffusion is kept fixed. For the same reason, the last block $[t_{m-2}, T]$ is updated by means of a Brownian motion $B_{[t_{m-2}, T]}(x_{t_{m-2}})$ starting at $X_{t_{m-2}} = x_{t_{m-2}}$, so that the value of the diffusion at T may vary. The acceptance coefficient of the move that update the last block is the same as in (12), with $[t_i, t_{i+2}] = [t_{m-2}, T]$ and $b_{[t_{m-2}, T]}(x_{t_{m-2}})$ in place of $bb_{[t_i, t_{i+2}]}(x_{t_i}, x_{t_{i+2}})$, where $b_{[t_{m-2}, T]}(x_{t_{m-2}})$ is the realization of the Brownian motion $B_{[t_{m-2}, T]}(x_{t_{m-2}})$.

This idea of updating smaller intervals at a time has been used in Shephard and Pitt (1997) for the simulation of non-Gaussian time series models, and later applied for the simulation of discretely observed diffusions, for example by Elerian, Chib, and Shephard (2001).

In Section 3.2 we shall illustrate the implementation of this algorithm by means of a toy example. Note that in this section and in the following we are considering base diffusions having drift linear in the parameter θ just for purposes of exposition.

3.2 Implementation of the algorithm: a toy example

We show here the implementation of the algorithm described in Section 3.1, by means of a toy example. Consider the model based on the diffusion process satisfying the SDE

$$dX_t = \theta_1 \sin(X_t)dt + \theta_2 dt + dB_t, \quad t \geq 0, \quad X_0 = 2, \quad (13)$$

with hazard function $h(u) = u^2$. We simulate observations from this model, for values of the parameters $\theta_1 = -1.4$ and $\theta_2 = -1$, and censoring time $C = 0.9$. In particular, we sample one realization x of the diffusion process satisfying (13), with $\theta_1 = -1.4$ and $\theta_2 = -1$. Then we simulate 200 i.i.d. observations from the corresponding distribution $F_{x,h} = 1 - \exp\left\{-\int_0^t (x_s)^2 ds\right\}$ and we censor the observations at a common cut-off $C = 0.9$. The diffusion is sampled at intervals of length 0.01, using Euler-Maruyama approximation. Figure 1 shows the corresponding hazards (the squared diffusion) and an histogram of sampled data. The hazard function has a typical shape, first (mainly) increasing and then (mainly) decreasing.

We choose as time horizon of interest $T = 1$. We then run the Hastings-within-Gibbs algorithm under the following specifications. The prior for (θ_1, θ_2) is Gaussian, as in Section 3.1, with $\mu_1 = -1.4$, $\mu_2 = -1$, $\lambda_{11} = \lambda_{22} = 1/5$ and $\lambda_{12} = 0$. The starting values of the parameters are $\theta_1 = \theta_2 = 0$, and the starting diffusion is a Brownian motion, starting at $x_0 = 2$. The diffusion path is updated on subintervals of length 0.2 at a time. The algorithm is run for 200000 iterations and the first 2000 are discarded as burn in.

Figure 2 shows the estimates of survival distribution, density, and hazard function, based on the MCMC output, together with pointwise approximate 90% highest posterior bands. The true survival distribution and hazard function are also displayed to evidence the good fit of the MCMC estimates. Figure 2 also shows autocorrelation functions for θ_1 and θ_2 series.

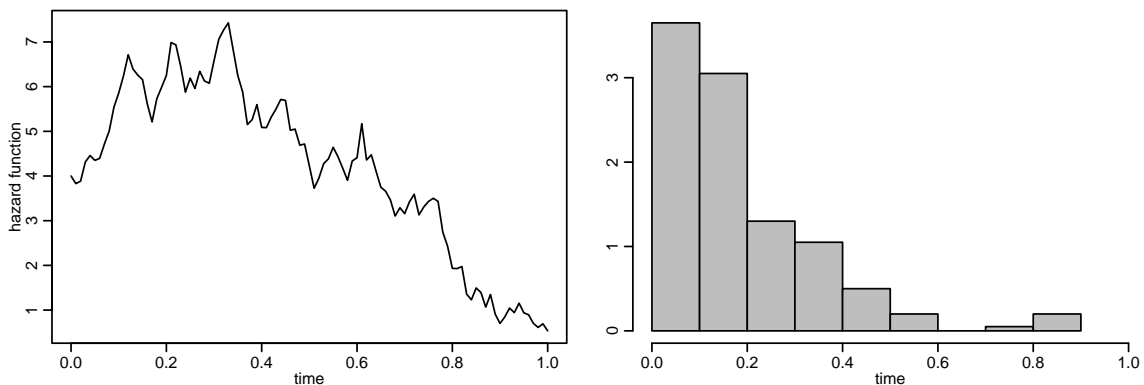


Figure 1: Left: hazard function x^2 . Right: histogram of data sampled from F_{x,x^2} with censoring at $C = 0.9$.

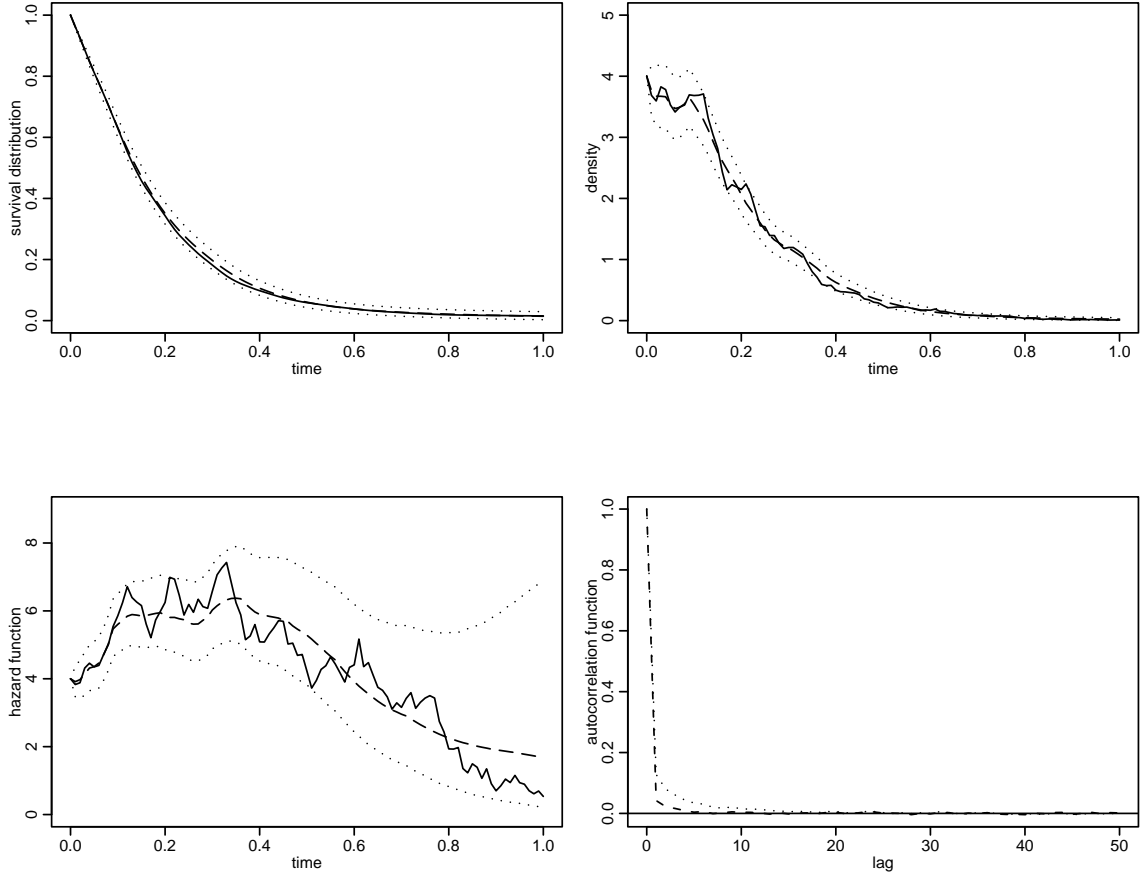


Figure 2: Top left: true survival distribution $1 - F_{x,x^2}$ (solid), together with its posterior mean (dashed) and pointwise approximate 90% highest posterior bands (dotted). Top right: true density (solid), together with its posterior mean (dashed) and pointwise approximate 90% highest posterior bands (dotted). Bottom left: true hazard function x^2 (solid), together with its posterior mean (dashed) and pointwise approximate 90% highest posterior bands (dotted). Bottom right: autocorrelation functions for θ_1 series (dotted) and θ_2 series (dashed).

4 Reparametrizations of the latent diffusion models

The MCMC algorithm described in the previous sections might have poor mixing properties when we consider a finite time horizon T significantly bigger than the maximum of the data. This problem is evidenced in figure 3. This figure shows the histogram of 200 i.i.d. observations from the distribution $F_{x',h}$, where x' is a new realization of the diffusion process satisfying the same SDE used in Section 3.2, and also the hazard function h and the censoring time C are the same. In this simulation we have fixed a longer time horizon $T = 1.8$, and we have then run the algorithm under the same specifications of Section 3.2. Figure 3 displays autocorrelation functions for θ_1 and θ_2 series, which are not exponentially decreasing. With the same dataset, but choosing a shorter time horizon (such as $T = 1$, as in the previous section), the algorithm does not exhibit strong serial correlation in the draws of θ_1 and θ_2 . The worsening of the mixing properties of the algorithm, when T becomes significantly bigger than the maximum of the data, was also observed for the dataset simulated in Section 3.2.

To avoid this problem, we propose a modification of the algorithm, which has good mixing

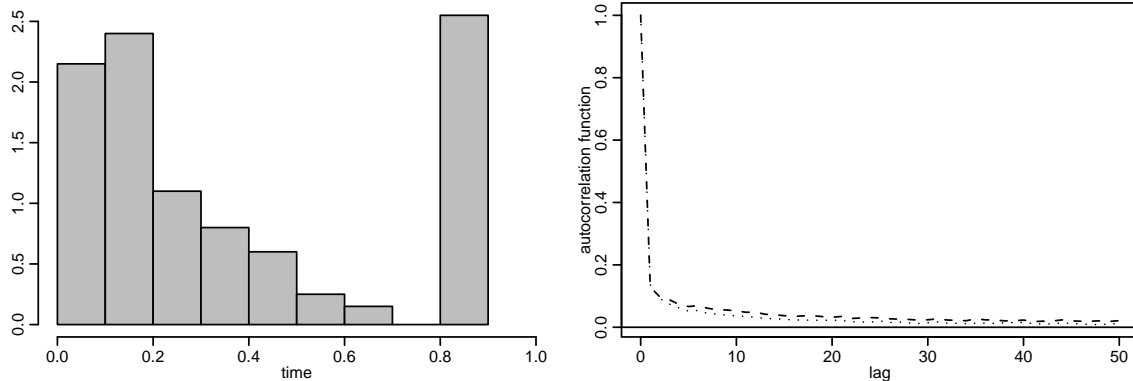


Figure 3: Right: histogram of data sampled from $F_{x',h}$ with censoring at $C = 0.9$. Right: autocorrelation functions for θ_1 series (dotted) and θ_2 series (dashed).

properties regardless of the choice of time horizon, and is in fact completely robust with respect to T . The algorithm is based on a simple reparametrization of the model. Indeed, the performance of MCMC methods, particularly when using Gibbs samplers, depends crucially on the parametrization of the unknown quantities in the hierarchical structure. The issue of reparametrization of the posterior distributions, as to improve convergence properties of the algorithms, has received much attention. See for example Hills and Smith (1992), Gelfand, Sahu, and Carlin (1995), Gelfand, Sahu, and Carlin (1996), and Papaspiliopoulos, Roberts, and Sköld (2003, 2007).

Instead of using the natural parametrization of the model in terms of (Θ, X) , the so-called *centered parametrization*, we parametrize it in terms of (Θ, \tilde{X}) , where

$$\tilde{X}_t = 1(t \leq y_{[n]}) X_t + 1(t > y_{[n]}) [B_t - B_{y_{[n]}}], \quad t \geq 0.$$

In the terminology used by Papaspiliopoulos, Roberts, and Sköld (2003), this is called a *partially non-centered parametrization*, the fully *non-centered parametrization* being, in this case, (Θ, B) . The diffusion X can then be reconstructed as function of Θ , \tilde{X} and y_1, \dots, y_n , by

$$\begin{cases} X_t = \tilde{X}_t & 0 \leq t \leq y_{[n]} \\ dX_t = \beta(X_t, \Theta)dt + \sigma d\tilde{X}_t & t \geq y_{[n]}. \end{cases}$$

The joint posterior distribution of Θ and \tilde{X} has density, with respect to the product measure $\mathcal{L}^d \otimes \mathbb{W}_\sigma$, given by

$$\pi(\theta, \tilde{x} | y_1, \dots, y_n) = C p_\Theta(\theta) g(x_{[0, y_{[n]}} | \theta) l(y_1, \dots, y_n | x_{[0, y_{[n]}}) \quad (14)$$

where $x_{[0, y_{[n]}} \equiv \tilde{x}_{[0, y_{[n]}}$, C is a normalizing constant, and $g(x_{[0, y_{[n]}} | \theta) = \frac{d\mathbb{P}_{y_{[n]}, \theta}}{d\mathbb{W}_{y_{[n]}, \sigma}}(x_{[0, y_{[n]}})$ is given by Girsanov's formula (2). Note in particular that (14) characterizes the posterior distribution of \tilde{X} , and thus the posterior distribution of the diffusion X , over the whole positive half-line. It thus also highlights that $X_{[0, y_{[n]}}$ acts as a sufficient statistics.

It is possible to simulate from (14) by means of a Gibbs sampler quite similar to the one described in Section 3.1. However, the algorithm is now completely robust to the choice of T , since the update of the parameter Θ , conditionally on \tilde{X} , only involve $\tilde{X}_{[0, y_{[n]}}$. In the first step, in fact, we now simulate Θ conditionally on $\tilde{X}_{[0, y_{[n]}}$. In the second step, we simulate \tilde{X} over the

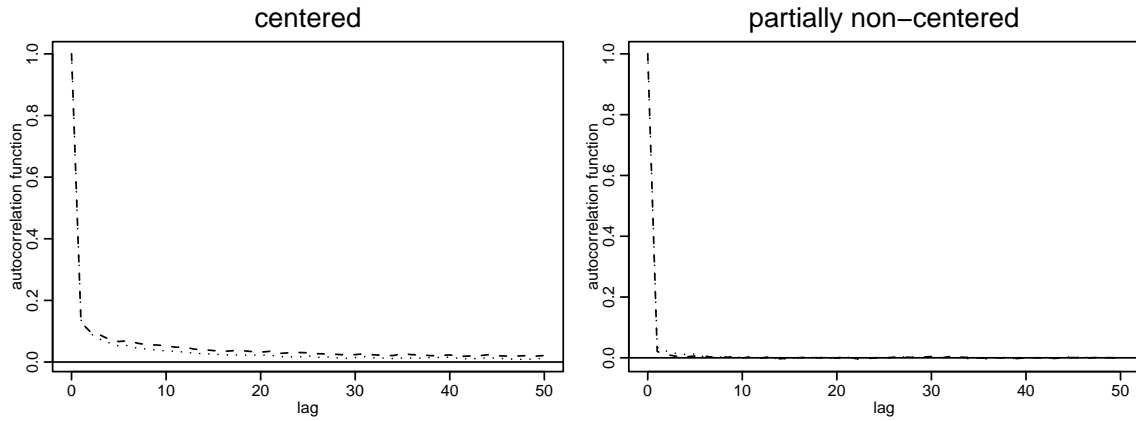


Figure 4: Autocorrelation functions for θ_1 series (dotted) and θ_2 series (dashed), obtained with the algorithm based on the centered parametrization (left) and with the algorithm based on the partially non-centered parametrization (right).

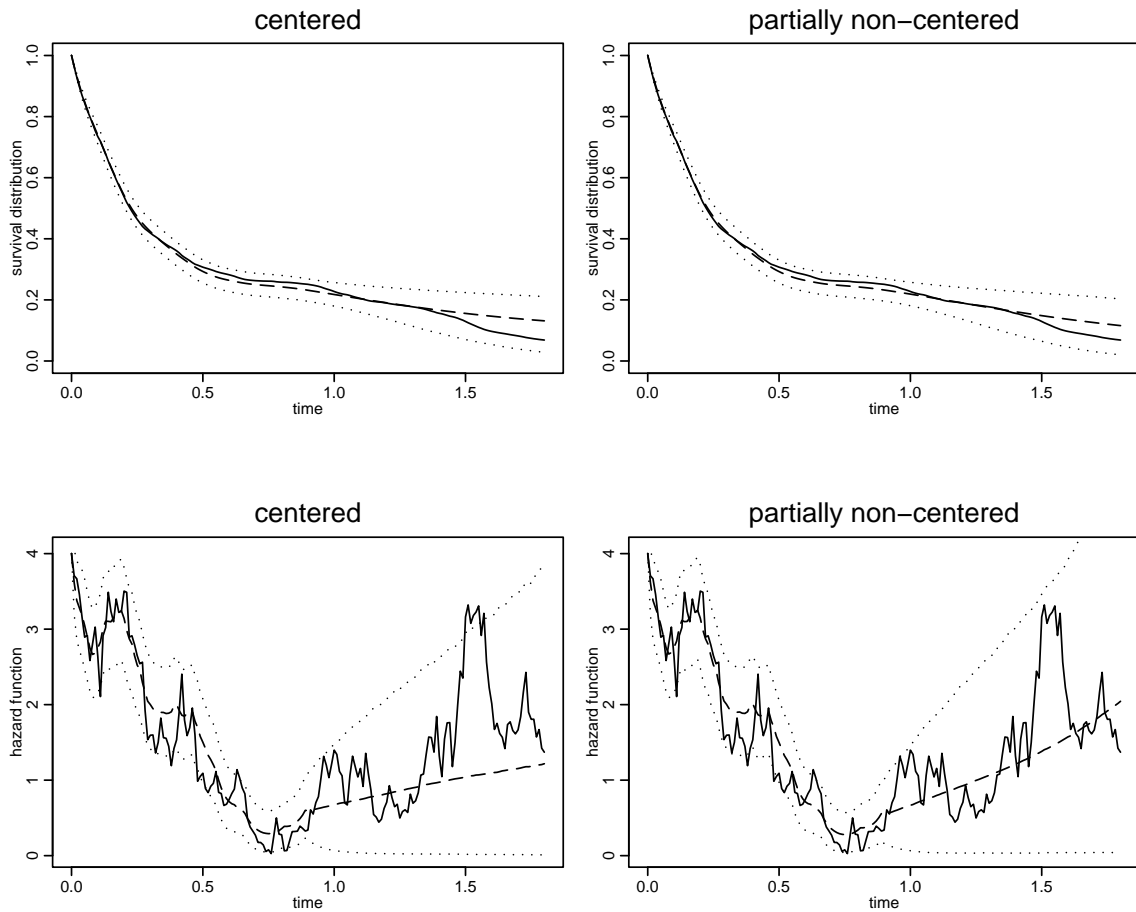


Figure 5: Top: true survival distribution $1 - F_{x',x,2}$ (solid), together with its posterior mean (dashed) and pointwise approximate 90% highest posterior bands (dotted), obtained with the algorithm based on the centered parametrization (left) and with the algorithm based on the partially non-centered parametrization (right). Bottom: true hazard function x'^2 (solid), together with its posterior mean (dashed) and pointwise approximate 90% highest posterior bands (dotted), obtained with the algorithm based on the centered parametrization (left) and with the algorithm based on the partially non-centered parametrization (right).

time interval of interest $[0, T]$, conditionally on Θ and the observations. In this case we use a proposal distribution which is a Brownian motion starting at x_0 , over the time interval $[0, y_{[n]}]$, and a Brownian motion starting at 0, over the time interval $[y_{[n]}, T]$. On $[0, y_{[n]}]$ we follow again the updating strategy, with the overlapping Brownian bridges, described in Section 3.1. When reconstructing the diffusion $X_{[0, T]}$, from Θ and $\tilde{X}_{[0, T]}$, we are careful to preserve the continuity of the diffusion path at time $y_{[n]}$. Details are omitted.

Figures 4 and 5 compares mixing and MCMC estimates obtained with the algorithms based on the centered parametrization and on the partially non-centered parametrization, for the dataset corresponding to figure 3. The specifications of the two algorithms are as in Section 3.2. Note that the hazard function is bathtub shaped. Hazard functions with such shape are quite common in survival analysis (think, for instance, to human mortality).

As we shall see in Section 6, another reparametrization of the model, that turns out to be useful in presence of covariates, is the fully *non-centered parametrization* in terms of (Θ, B) . The diffusion X can be reconstructed as function of Θ and B , simply by the SDE

$$dX_t = \beta(X_t, \Theta)dt + \sigma dB_t, \quad t \geq 0, \quad X_0 = x_0.$$

The joint posterior distribution of Θ and B has density, with respect to the product measure $\mathcal{L}^d \otimes \mathbb{W}_\sigma$, given by

$$\pi(\theta, b | y_1, \dots, y_n) = C p_\Theta(\theta) l(y_1, \dots, y_n | \theta, b_{[0, y_{[n]}]}) \quad (15)$$

where C is a normalizing constant, and $l(y_1, \dots, y_n | \theta, b_{[0, y_{[n]}]}) = l(y_1, \dots, y_n | x_{[0, y_{[n]}]})$ is as in (4). Note that, similarly to what has been noticed for the partially non-centered parametrization, also (15) characterizes the posterior distribution of the diffusion X over the whole positive half-line. Moreover, also in this case the Gibbs sampler that simulate from (15) is completely robust with respect to the choice of the time horizon T . In the first step, we simulate Θ conditionally on $B_{[0, y_{[n]}]}$ and the observations. Note in particular that the conditional distribution of Θ , given $B_{[0, T]}$ and the observations, has now density, with respect to \mathcal{L}^d , proportional to $p_\Theta(\theta) l(y_1, \dots, y_n | \theta, b_{[0, y_{[n]}]})$. In the second step, we simulate B over the time interval of interest $[0, T]$, conditionally on Θ and the observations. We use for proposal distribution a Brownian motion starting at 0, and we employ the updating strategy based on overlapping Brownian bridges. In this case, when updating the Brownian motion path b over the subinterval $[t_i, t_{i+2}]$, we need to reconstruct the corresponding diffusion path x over the subinterval $[t_i, T]$, in order to preserve the continuity of the diffusion path at time t_{i+2} . Details are omitted.

5 Latent diffusion models for multiple groups of observations

We now discuss a straightforward generalization of the framework developed in the previous sections, and deal with the case of multiple groups of observations, where the observations within each group are taken under homogeneous conditions. Consider for example the case in which different treatments are being administered to different groups of patients in a clinical trial.

Given $\Theta = \theta$, let $X^{[1]}, \dots, X^{[q]}$ be q stochastically independent diffusion processes satisfying (1), and $F_{X^{[1]}, h}, \dots, F_{X^{[q]}, h}$ the relative random distributions as in (3). Now consider q sequences

of observations $(Y_n^{[1]})_n, \dots, (Y_n^{[q]})_n$ such that the random variables in $((Y_n^{[1]})_n, \dots, (Y_n^{[q]})_n)$ are conditionally independent, given $F_{X^{[1]}, h}, \dots, F_{X^{[q]}, h}$, and the random variables in $(Y_n^{[k]})_n$ have common distribution $F_{X^{[k]}, h}$, for $k = 1, \dots, q$.

The joint distribution of $Y_1^{[1]}, \dots, Y_{n_1}^{[1]}, \dots, Y_1^{[q]}, \dots, Y_{n_q}^{[q]}$, given $X^{[1]} = x^{[1]}, \dots, X^{[q]} = x^{[q]}$, has density, with respect to \mathcal{L}^n (where $n = n_1 + \dots + n_q$), given by

$$l(y_1^{[1]}, \dots, y_{n_1}^{[1]}; \dots; y_1^{[q]}, \dots, y_{n_q}^{[q]} | x_{[0, y_{[n_1]}]}^{[1]}, \dots, x_{[0, y_{[n_q]}]}^{[q]}) = \prod_{k=1}^q l(y_1^{[k]}, \dots, y_{n_k}^{[k]} | x_{[0, y_{[n_k]}]}^{[k]})$$

where $y_{[n_k]} := \max\{y_1^{[k]}, \dots, y_{n_k}^{[k]}\}$ and $l(y_1^{[k]}, \dots, y_{n_k}^{[k]} | x_{[0, y_{[n_k]}]}^{[k]})$ is as in (4). Using the partially non-centered parametrization described in Section 4, the joint posterior distribution of Θ and $\tilde{X}^{[1]}, \dots, \tilde{X}^{[q]}$ has density, with respect to the product measure $\mathcal{L}^d \otimes \mathbb{W}_\sigma^q$, given by

$$\begin{aligned} & \pi(\theta, \tilde{x}^{[1]}, \dots, \tilde{x}^{[q]} | y_1^{[1]}, \dots, y_{n_1}^{[1]}; \dots; y_1^{[q]}, \dots, y_{n_q}^{[q]}) \\ &= C p_\Theta(\theta) \left[\prod_{k=1}^q g(x_{[0, y_{[n_k]}]}^{[k]} | \theta) l(y_1^{[k]}, \dots, y_{n_k}^{[k]} | x_{[0, y_{[n_k]}]}^{[k]}) \right] \end{aligned} \quad (16)$$

where C is a normalizing constant, and $g(x_{[0, y_{[n_k]}]}^{[k]} | \theta) = \frac{d\mathbb{P}_{y_{[n_k]}^{[k]}, \theta}}{d\mathbb{W}_{y_{[n_k]}^{[k]}, \sigma}}(x_{[0, y_{[n_k]}]}^{[k]})$ is given by Girsanov's formula (2).

The contributions of the q groups of observations factorize in (16), and a simple modification of the MCMC algorithm presented in the previous sections may be used to deal with this case. Let T_1, \dots, T_q be the time horizons of interest for the q groups, with $T_k \geq y_{[n_k]}$ for $k = 1, \dots, q$. The Hastings-within-Gibbs algorithm for sampling from (16) alternates between

1. simulation of Θ , conditional on the current paths of $\tilde{X}_{[0, y_{[n_1]}]}^{[1]}, \dots, \tilde{X}_{[0, y_{[n_q]}]}^{[q]}$;
2. for each k in $\{1, \dots, q\}$, simulation of $\tilde{X}_{[0, T_k]}^{[k]}$, conditional on the observations $Y_1^{[k]}, \dots, Y_{n_k}^{[k]}$, and the current value of Θ .

Consider, for example, a latent diffusion model with q stochastically independent diffusion processes, $X^{[1]}, \dots, X^{[q]}$, satisfying the SDE (10). Choose the same multivariate Gaussian prior for Θ that has been used in Section 3.1. Then, the distribution of Θ , given $\tilde{X}_{[0, y_{[n_1]}]}^{[1]} = x_{[0, y_{[n_1]}]}^{[1]}, \dots, \tilde{X}_{[0, y_{[n_q]}]}^{[q]} = x_{[0, y_{[n_q]}]}^{[q]}$, is still Gaussian, with mean vector and covariance matrix as in (11), but with

$$S_i := \frac{1}{\sigma^2} \left[\sum_{k=1}^q \int_0^{y_{[n_k]}} f_i(x_t^{[k]}) dx_t^{[k]} \right] + \sum_{j=1}^d \lambda_{ij} \mu_j \quad L_{ij} := \frac{1}{\sigma^2} \left[\sum_{k=1}^q \int_0^{y_{[n_k]}} f_i(x_t^{[k]}) f_j(x_t^{[k]}) dt \right] + \lambda_{ij}$$

for $i = 1, \dots, d$, $j = 1, \dots, d$. The update of the parameter Θ can thus be performed by sampling directly from this conditional distribution. The second step may be carried out by q repetitions of the updating mechanism described in Sections 3.1 and 4.

Note that we are here considering a simple hierarchical structure, where inference on the separate groups is linked only at the level of the finite dimensional parameter Θ . For some applications this might allow too little borrowing of strength for inference across groups of patients. In Section 6 we shall instead describe a more complex hierarchical structure, suitable in the presence of covariates and that allows for a much stronger borrowing of strength for inference across individuals.

5.1 An illustrative application to a real dataset with multiple groups of observations

In this section we show the implementation of the latent diffusion model for multiple groups of observations via an illustrative application to a small dataset from a clinical trial, that has been considered in a number of papers in the context of survival analysis, among which Gehan (1965), Cox (1972), Wei (1984) and Xu and O’Quigley (2000) in the non-Bayesian literature, and Kalbfleisch (1978), Laud, Damien, and Smith (1998) and Damien and Walker (2002) in the Bayesian one. In the trial, reported by Freireich (1963), 6-mercaptopurine (6-MP) was compared to a placebo in the maintenance of remission in acute leukemia. The following lengths of remission in weeks were recorded for 42 patients, half of which treated with the 6-MP drug and half with the placebo (a + sign indicates a censored observation):

6-MP: 6, 6, 6, 6+, 7, 9+, 10, 10+, 11+, 13, 16, 17+, 19+, 20+, 22, 23, 25+, 32+, 32+, 34+, 35+
 placebo: 1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23.

We thus consider a model for two groups of observations, namely the 6-MP drug group and the placebo group. As latent diffusion model, we shall use the stochastic perturbation around the Weibull, described in Section 2.1; recall that this model has base diffusion satisfying the SDE

$$dX_t = \theta_1 (\text{sign}(X_t)) |X_t|^{\theta_2} dt + \sigma dB_t, \quad t \geq 0, \quad X_0 = x_0 > 0,$$

and hazard function $h(u) = |u|$.

We express the data as fractions of one year, and choose as time horizons of interest $T_1 = T_2 = 0.75$, corresponding to 9 months (39 weeks). We take Θ_1 and Θ_2 a priori independent, with a Gaussian prior distribution for Θ_1 , with mean $\mu = 0$ and variance $1/\lambda = 5$, and a uniform prior over $[0, 1]$ for Θ_2 . We moreover set $x_0 = 0.8$ and $\sigma = 8$. We then run the Hastings-within-Gibbs algorithm based on the partially non-centered parametrization. The update of Θ_1 is performed by sampling directly from the conditional distribution Θ_1 given $\Theta_2, \tilde{X}_{[0, y_{[n_1]}]}^{[1]}, \tilde{X}_{[0, y_{[n_2]}]}^{[2]}$, which is still Gaussian with mean $\frac{S + \lambda \mu}{L + \lambda}$ and variance $\frac{1}{L + \lambda}$, where

$$S := \frac{1}{\sigma^2} \left[\sum_{j=1}^2 \int_0^{y_{[n_j]}} ((\text{sign}(x_t^{[j]})) |x_t^{[j]}|^{\theta_2}) dx_t^{[j]} \right] \quad L := \frac{1}{\sigma^2} \left[\sum_{j=1}^2 \int_0^{y_{[n_j]}} (|x_t^{[j]}|^{\theta_2})^2 dt \right]$$

For the update of Θ_2 we use an independence sampler with a Beta proposal distribution, with parameters $(1/2, 1/2)$. The update of $\tilde{X}^{[1]}$ and $\tilde{X}^{[2]}$ is carried out as described in the previous sections. The algorithm is run for 200000 iterations and the first 2000 are discarded as burn in.

Figure 6 displays the MCMC estimates of the survival distributions of the two groups, 6-MP drug and placebo, together with the relative Kaplan-Meier curves. Note that the MCMC estimates of the two survival distributions are closer one another than the two Kaplan-Meier curves, thus showing borrowing of strength for inference among the two groups. Hence, the latent diffusion model, that gains much flexibility over a fully parametric model by introducing randomness around it, does not suffer from the opposite problem of being too data-driven. Figure 6 also displays the MCMC estimates of the hazards of the two groups.

We could now verify the efficacy of 6-MP drug treatment as proposed in Damien and Walker (2002). In particular, under the hypothesis that 6-MP drug is inefficient, we would regard all

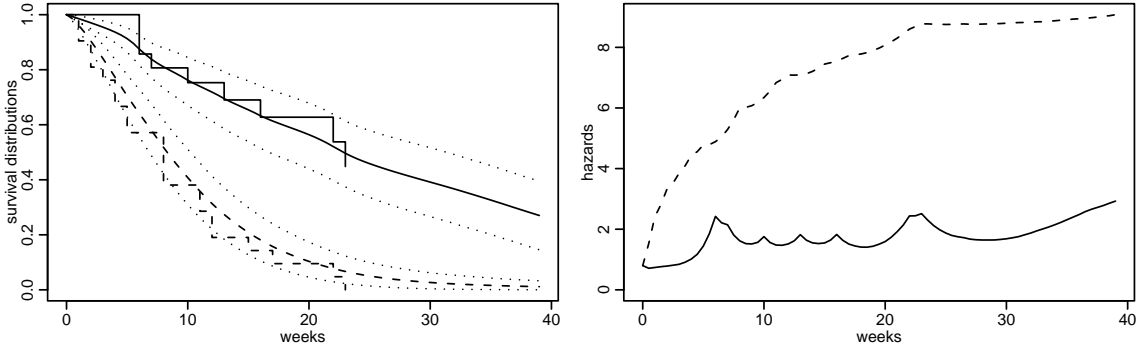


Figure 6: Left: posterior mean survival distributions and pointwise approximate 90% highest posterior bands, for the group of patients treated with 6-MP drug (solid) and for the group of patients treated with the placebo (dashed), together with corresponding Kaplan-Meier curves. Right: posterior mean hazards for the group of patients treated with 6-MP drug (solid) and for the group of patients treated with the placebo (dashed).

patients as belonging to 1 single group, instead of 2. We could then implement the latent diffusion model based on the stochastic perturbation of the Weibull, but with just 1 diffusion process. Call M_1 the model where all patients belong to 1 single group (corresponding to the hypothesis H_1 of null efficacy of 6-MP drug), and call M_2 the model considered above (corresponding to the hypothesis H_2 of efficacy of 6-MP drug). If the a priori probabilities of hypothesis H_1 and H_2 are set equal to 0.5, the Bayes Factor

$$\text{BF} = \frac{\text{probability density of data under model } M_1}{\text{probability density of data under model } M_2}$$

gives the posterior odds in favor of H_1 . As expected, the computed Bayes Factor ($\text{BF} = 9 \times 10^{-6}$) gives a strong evidence of the efficacy of 6-MP drug.

6 Latent diffusion models with covariates

Covariates can be included in the latent diffusion models described in a very natural way, as influencing directly the underlying diffusion. For instance, if \mathbf{Z} is a vector of p covariates measured at time 0, we can use the model based on the diffusion satisfying the SDE

$$\begin{aligned} dX_t &= \beta(X_t, \mathbf{z}, \theta) dt + \sigma dB_t & t \geq 0 \\ X_0 &= x_0(\mathbf{z}, \theta). \end{aligned} \tag{17}$$

In particular, following suggestions of Aalen and Gjessing (2001) and Aalen et al. (2008) for barrier hitting models, those covariates which represent measures of how far the underlying process, that leads to the event, has advanced (such as staging measures in cancer) may be taken to influence the starting point of the diffusion. Those covariates which instead represent causal influence on the development of the process may be taken to influence the drift of the diffusion.

Let \mathbf{z} take values $\mathbf{z}^{[1]}, \dots, \mathbf{z}^{[a]}$. Then (17) gives q different diffusions, $X^{[\mathbf{z}=\mathbf{z}^{[1]}]}, \dots, X^{[\mathbf{z}=\mathbf{z}^{[a]}]}$, driven by the same Brownian motion B , with

$$\begin{aligned} dX_t^{[\mathbf{z}=\mathbf{z}^{[k]}]} &= \beta(X_t^{[\mathbf{z}=\mathbf{z}^{[k]}]}, \mathbf{z}^{[k]}, \theta) dt + \sigma dB_t & t \geq 0 \\ X_0 &= x_0(\mathbf{z}^{[k]}) \end{aligned}$$

for $k = 1, \dots, q$. Denote by $F_{X^{[z=z^{[1]}]_h}, \dots, F_{X^{[z=z^{[q]}]_h}}$ the relative random distributions as in (3). Moreover, denote by $Y_1^{[z=z^{[k]}]}, \dots, Y_{n_k}^{[z=z^{[k]}]}$ the survival times of the n_k individuals having covariates $\mathbf{z} = \mathbf{z}^{[k]}$, for $k = 1, \dots, q$. The survival times $Y_1^{[z=z^{[1]}]}, \dots, Y_{n_k}^{[z=z^{[k]}]}$, conditionally on $F_{X^{[z=z^{[k]}]_h}}$, are i.i.d. with common distribution $F_{X^{[z=z^{[k]}]_h}}$. Since the q diffusions are driven by the same Brownian motion, it is here more natural to use the fully non-centered parametrization of the model, described in Section 4. In particular, the joint distribution of $Y_1^{[z=z^{[1]}]}, \dots, Y_{n_1}^{[z=z^{[1]}]}, \dots, Y_1^{[z=z^{[q]}]}, \dots, Y_{n_q}^{[z=z^{[q]}]}$, given $B = b$ and $\Theta = \theta$, has density, with respect to \mathcal{L}^n (where $n = n_1 + \dots + n_q$), given by

$$\begin{aligned} & l(y_1^{[z=z^{[1]}]}, \dots, y_{n_1}^{[z=z^{[1]}]}; \dots; y_1^{[z=z^{[q]}]}, \dots, y_{n_q}^{[z=z^{[q]}]} | \theta, b_{[0, y_{[n]}]}, \mathbf{z}^{[1]}, \dots, \mathbf{z}^{[q]}) = \\ & = \prod_{k=1}^q l(y_1^{[z=z^{[k]}]}, \dots, y_{n_k}^{[z=z^{[k]}]} | \theta, b_{[0, y_{[n_k]}]}, \mathbf{z}^{[k]}) \end{aligned}$$

where $y_{[n]} := \max\{y_1, \dots, y_n\}$, $y_{[n_k]} := \max\{y_1^{[z=z^{[k]}]}, \dots, y_{n_k}^{[z=z^{[k]}]}\}$, and

$$l(y_1^{[z=z^{[k]}]}, \dots, y_{n_k}^{[z=z^{[k]}]} | \theta, b_{[0, y_{[n_k]}]}, \mathbf{z}^{[k]}) = l(y_1^{[z=z^{[k]}]}, \dots, y_{n_k}^{[z=z^{[k]}]} | x_{[0, y_{[n_k]}]}^{[z=z^{[k]}]})$$

is as in (4). The joint posterior distribution of Θ and B has density, with respect to the product measure $\mathcal{L}^d \otimes \mathbb{W}_\sigma$, given by

$$\begin{aligned} & \pi(\theta, b | y_1^{[z=z^{[1]}]}, \dots, y_{n_1}^{[z=z^{[1]}]}; \dots; y_1^{[z=z^{[q]}]}, \dots, y_{n_q}^{[z=z^{[q]}]}; \mathbf{z}^{[1]}, \dots, \mathbf{z}^{[q]}) \\ & = C p_\Theta(\theta) \prod_{k=1}^q l(y_1^{[z=z^{[k]}]}, \dots, y_{n_k}^{[z=z^{[k]}]} | \theta, b_{[0, y_{[n_k]}]}, \mathbf{z}^{[k]}). \end{aligned} \quad (18)$$

Note that this model is structurally different from the model for multiple groups of observations, described in Section 5, since the distributions of the survival times are here linked at the level of the Brownian motion, allowing a much stronger borrowing of strength for inference across individuals who share a common value of even just one of the p covariates.

Denote as usual by T the time horizon of interest, $T \geq y_{[n]}$. The Hastings-within-Gibbs algorithm for sampling from (18) alternates between

1. simulation of Θ , conditional on the current path of $B_{[0, y_{[n]}]}$, the observations and the covariates;
2. simulation of $B_{[0, T]}$, conditional on the current value of Θ , the observations and the covariates.

In particular, the update of the Brownian motion $B_{[0, T]}$ can be carried out via the updating strategy based on overlapping Brownian bridges, as described in Section 4.

6.1 An illustrative application to a real dataset with covariates

In this section we illustrate how to efficiently handle the model with covariates, via an application to a dataset concerning 272 patients diagnosed with non-small cell lung cancer. The dataset is described in detail by Muers et al. (1996). Survival times are measured in months from the time of diagnosis (with 17% of censoring) and some covariates are recorded at the time of diagnosis. Just to give an illustration of the model, we shall consider here two covariates: sex (F=0: male and F=1: female) and hoarseness (H=0: absent and H=1: present). Using for instance the model based on the stochastic perturbation around the Weibull, we can include these covariates as follows:

$$\begin{aligned} dX_t &= \exp\{\theta_{10} + \theta_{11} F\} (\text{sign}(X_t)) |X_t|^{\theta_2} dt + \sigma dB_t \quad t \geq 0 \\ X_0 &= \exp\{\theta_{00} + \theta_{01} F + \theta_{02} H\} \end{aligned}$$

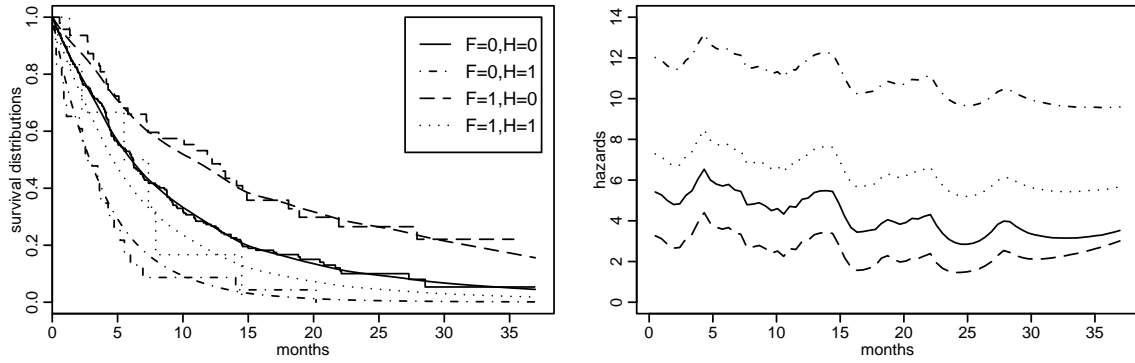


Figure 7: Left: posterior mean survival distributions together with Kaplan-Meier curves, for male patients without hoarseness at time of diagnosis ($F = 0, H = 0$, solid line), for male patients with hoarseness ($F = 0, H = 1$, dot and dash line), for female patients without hoarseness ($F = 1, H = 0$, dashed line), and for female patients with hoarseness ($F = 1, H = 1$, dotted line). Right: the same for posterior mean hazard functions.

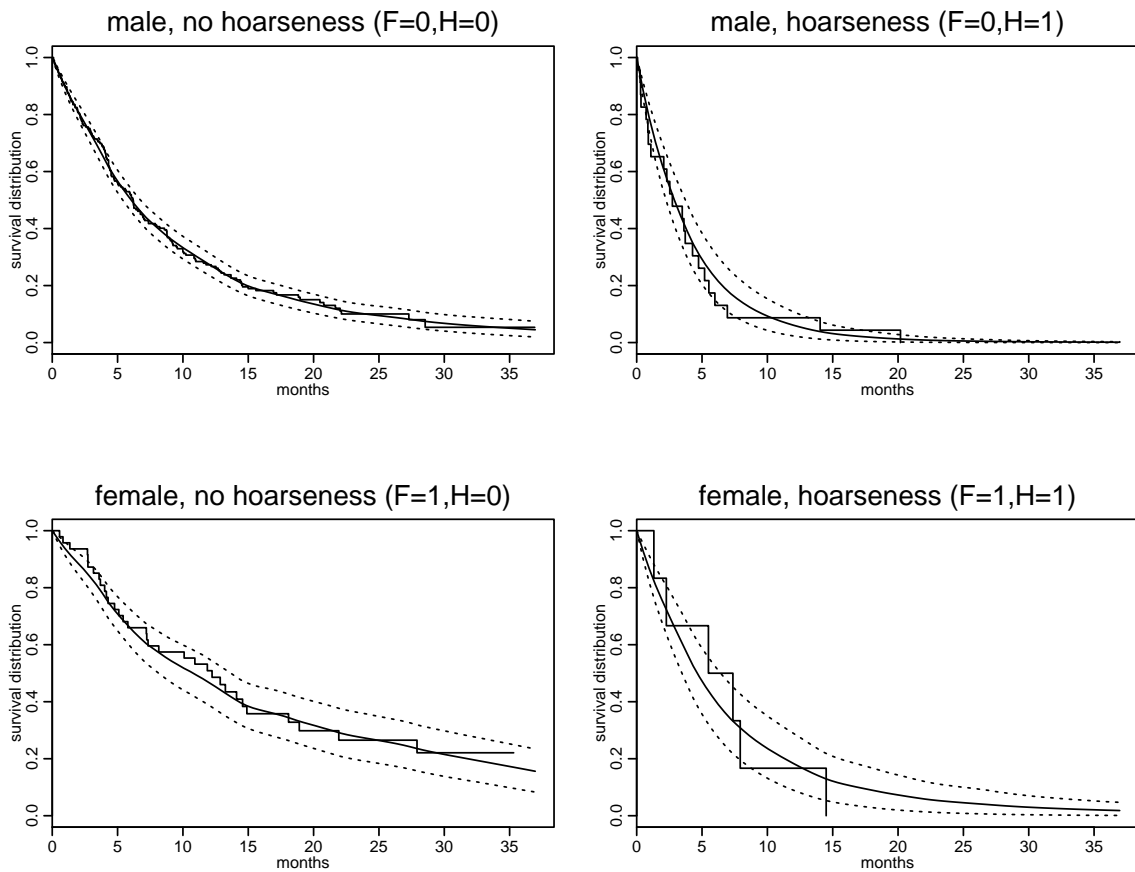


Figure 8: Top left: posterior mean survival distribution and pointwise approximate 90% highest posterior bands, together with Kaplan-Meier curve, for male patients without hoarseness. Top right: the same for male patients with hoarseness. Bottom left: the same for female patients without hoarseness. Bottom right: the same for female patients with hoarseness.

Note that, following the suggestion of Aalen et al. (2008), we have modeled the covariate hoarseness, which only represents a measure of how far the lung tumor has advanced, as influencing the starting point of the diffusion; we have instead taken the covariate sex to influence both the starting point and the drift of the diffusion, in order to account for possible differences between males and females both in the hazards at time of diagnosis and in the hazard dynamics. The covariates combinations determine four different diffusions, $X^{[F=0,H=0]}$, $X^{[F=0,H=1]}$, $X^{[F=1,H=0]}$ and $X^{[F=1,H=1]}$, driven by the same Brownian motion. According to this model, the hazard at time 0 (the time of diagnosis) of patients suffering of hoarseness is $\exp\{\theta_{02}\}$ times the one of patients not suffering of hoarseness, and the hazard at time 0 of female patients is $\exp\{\theta_{01}\}$ times the one of male patients; moreover, $\exp\{\theta_{11}\}$ gives a measure of the different progression rate of the cancer in female patients with respect to male patients.

We express the data as fractions of a quadriennium, and choose as time horizon T the maximum of the observations, corresponding to about 37 months. In order to avoid dependencies among the $(\theta_{00}, \theta_{01}, \theta_{02})$ -parameters and among the $(\theta_{10}, \theta_{11})$ -parameters, we reparametrize them in terms of $(\eta_{00}, \theta_{01}, \theta_{02})$ and (η_{10}, θ_{11}) , with $\theta_{00} = \eta_{00} - p_F \theta_{01} - p_H \theta_{02}$ and $\theta_{10} = \eta_{10} - p_F \theta_{11}$, where we have denoted by p_F and p_H the percentage of females patients and the percentage of patients suffering of hoarseness, respectively. We take all the parameters to be a priori independent, with Gaussian priors with mean 0 and variance 5 for all the parameters but Θ_2 , for which we use a uniform prior over $[0, 1]$. We moreover set $\sigma = 8$. We then run the Hastings-within-Gibbs algorithm based on the non-centered parametrization of the model. The update of the parameters is performed via independence samplers having proposal distributions equal to the priors. The algorithm is run for 200000 iterations and the first 2000 are discarded as burn in.

Figure 7 shows posterior mean survival distributions together with Kaplan-Meier curves, for male patients without hoarseness at time of diagnosis ($F = 0, H = 0$, solid line), for male patients with hoarseness ($F = 0, H = 1$, dot and dash line), for female patients without hoarseness ($F = 1, H = 0$, dashed line), and for female patients with hoarseness ($F = 1, H = 1$, dotted line). The four survivals are also plotted separately in Figure 8, with 90% highest posterior bands. Figure 7 also displays the posterior mean hazard functions for the four covariates combinations. In particular, the posterior mean hazard at time 0 of patients suffering of hoarseness is 2.2 times bigger than the one of patients not suffering of hoarseness; whereas, the hazard at time 0 of female patients is 0.6 times the one of male patients.

Note that, even if in this illustrative application we have only considered categorical covariates, also quantitative covariates can be included in the model; it may though be necessary to categorize these covariates, in order to have a sufficient number of observations for each of the diffusion processes. This of course requires larger datasets.

7 Generalization to the case of unknown diffusion coefficient

An important generalization of the models we have considered so far consists in considering diffusion processes with unknown diffusion coefficient σ , since σ describes a natural measure of prior uncertainty. We briefly discuss how to deal with this case.

Let Σ be a real random variable. Given $\Theta = \theta$ and $\Sigma = \sigma$, consider the scalar diffusion process X solution of the SDE (1), and denote by $\mathbb{P}_{T,\theta,\sigma}$ the law of $X_{[0,T]}$. Let $p_\Sigma(\cdot)$ be the prior

density, with respect to \mathcal{L} , of Σ (for simplicity, we take Θ and Σ to be stochastically independent a priori). Let us consider for instance the centered parametrization of the model. The joint posterior distribution of $(\Theta, \Sigma, X_{[0,T]})$ has density, with respect to $\mathcal{L}^{d+1} \otimes \mathbb{W}_{T,\sigma}$, given by

$$\pi(\theta, \sigma, x_{[0,T]} | y_1, \dots, y_n) = C p_{\Theta}(\theta) p_{\Sigma}(\sigma) g(x_{[0,T]} | \theta, \sigma) l(y_1, \dots, y_n | x_{[0,y_{[n]}]}) \quad (19)$$

where C is a normalizing constant, and $g(x_{[0,T]} | \theta, \sigma) := \frac{d\mathbb{P}_{T,\theta,\sigma}}{d\mathbb{W}_{T,\sigma}}(x_{[0,T]})$ is given by Girsanov's formula (2).

The quadratic variation of a diffusion processes, having diffusion coefficient σ , satisfies

$$\lim_{m \rightarrow \infty} \sum_{i=1}^m (X_{ti/m} - X_{t(i-1)/m})^2 = t\sigma^2 \quad \mathbb{W}_{T,\sigma} - \text{a.s. for all } t.$$

Therefore, the conditional distribution of Σ , given the diffusion $X_{[0,T]}$, degenerates to a point mass, and Σ is completely determined by the diffusion path. In practice, we cannot simulate the diffusion path in continuous time, but just at discrete time instants. Anyway, the finer the time discrete approximation $\{X_{iT/m} : i = 1, \dots, m\}$ of the diffusion $X_{[0,T]}$, the stronger becomes the dependence between $\{X_{iT/m} : i = 1, \dots, m\}$ and Σ . Consider the algorithm for the simulation from (19), that alternates between

1. simulation of Θ , conditional on the current value of Σ and the current path of $X_{[0,T]}$;
2. simulation of Σ , conditional on the current value of Θ and the current path of $X_{[0,T]}$;
3. simulation of $X_{[0,T]}$, conditional on the observations and the current values of Θ and Σ .

The finer the approximation of the diffusion path, the worse the convergence of the algorithm becomes. In the limiting case $m = \infty$ (that is, if the diffusion process could be simulated in continuous time), this scheme would be reducible. See Roberts and Stramer (2001). An alternative way to see this problem is to note that the collection of measures $\{\mathbb{W}_{T,\sigma} : \sigma \in \mathbb{R}\}$ are mutually singular, and therefore so are the measures $\{\mathbb{P}_{T,\theta,\sigma} : \sigma \in \mathbb{R}\}$.

In this case, the need for a different parametrization of the model is thus compelling. Following Roberts and Stramer (2001), we parametrize the model in terms of $(\Theta, \Sigma, \dot{X})$, where $\dot{X}_t = (X_t - X_0)/\Sigma$. By *Itô's formula*,

$$d\dot{X}_t = \frac{\beta(\dot{X}_t, \Theta)}{\Sigma} dt + dB_t, \quad t \geq 0, \quad \dot{X}_0 = 0.$$

The distribution of $\dot{X}_{[0,T]}$ depends on Σ , but any realization of $\dot{X}_{[0,T]}$ contains only finite information about Σ . Analogous reparametrizations are derived starting from the ones described in Section 4. MCMC algorithms based on these reparametrizations can be obtained as simple modifications of the ones previously described.

Consider the toy example described in Section 3.2, and assume the same model, but let the diffusion process have an unknown diffusion coefficient. Let the prior for this coefficient be exponential with mean 1. Figure 9 displays the results obtained with the MCMC algorithm based on the reparametrization $(\Theta, \Sigma, \dot{X})$. Specification of the algorithm are as in Section 3.2. Note that the mixing for σ is slow relatively to the very good mixing for θ_1 and θ_2 , but this does not prevent good estimates of the survival distribution, density and hazard being obtained. Slow mixing for σ could be probably improved by a further reparametrization of the model.

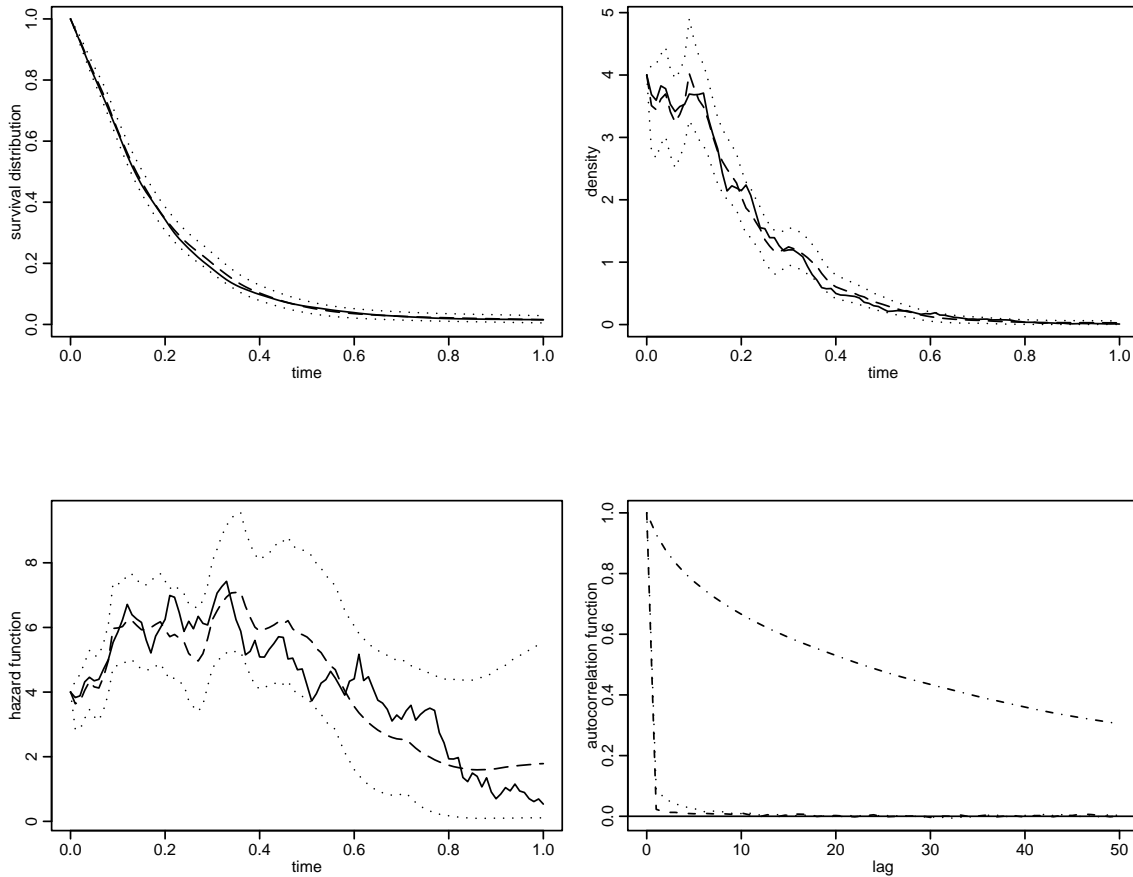


Figure 9: As in Figure (2), but for the model with unknown diffusion coefficient. Bottom right plot also displays autocorrelation function for σ series (dotdash line).

Alternatively to the case of unknown diffusion coefficient, it would be possible to consider models based on diffusion processes having $\sigma = 1$, but with hazard function $h(\Gamma, X)$, where Γ is a random parameter. Also in this case, a reparametrization of the model would be necessary.

8 Discussion

In this paper we have described latent diffusion models for survival analysis and we have shown that these models can be efficiently treated by means of MCMC techniques. We have dealt with the case of multiple groups of observations, typical of clinical trials, and we have shown how covariates can be efficiently included in the models. We have outlined how in the described framework it is possible to consider stochastic perturbations of common survival models. In particular, we have used a stochastic perturbation of the Weibull model in some illustrative applications to small datasets, with multiple groups of observations and with covariates. Applications to larger datasets, where the potential of latent diffusion model may be fully expressed, will be object of future work. All analyses presented are computationally feasible within R (see R Development Core Team (2007)).

Another generalization of the model we intend to explore regards random probabilities based on jump diffusion processes. As noticed in Section 2, the cumulative hazard functions, associated with

random probabilities based on diffusions, are smooth, being the integrals of continuous processes. By replacing the diffusion process with a jump diffusion process it would be possible to capture sudden changes in the behavior of cumulative hazards, that might be due to some kind of shock experienced by the population. Hazards modeled through stochastic processes with jumps have been studied for instance by Gjessing et al. (2003).

ACKNOWLEDGMENTS

We would like to thank Robin Henderson and Piercesare Secchi for useful comments, and Omiros Papaspiliopoulos and Alexandros Beskos for their help with the programming. We are also grateful to the associate editor and two anonymous referees for their constructive comments. The second author acknowledge funding by EC Marie Curie Training Site Human Potential Programme, to visit the Department of Mathematics and Statistics, Lancaster University, and by the Centre for Research in Statistical Methodology (CRiSM), University of Warwick.

References

- Aalen, O. O., Borgan, Ø., and Gjessing, H. K. (2008), *Survival and Event History Analysis. A process point of view*, Springer Series in Statistics for Biology and Health, New York: Springer.
- Aalen, O. O. and Gjessing, H. K. (2001), “Understanding the shape of the hazard rate: a process point of view,” *Statist. Sci.*, 16, 1–22, with comments and a rejoinder by the authors.
- (2004), “Survival models based on the Ornstein-Uhlenbeck process,” *Lifetime Data Anal.*, 10, 407–423.
- Beskos, A., Papaspiliopoulos, O., Roberts, G. O., and Fearnhead, P. (2006), “Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes,” *J. R. Statist. Soc. B*, to appear.
- Cox, D. R. (1972), “Regression models and life-tables,” *J. Roy. Statist. Soc. Ser. B*, 34, 187–220, with discussion by F. Downton, Richard Peto, D. J. Bartholomew, D. V. Lindley, P. W. Glassborow, D. E. Barton, Susannah Howard, B. Benjamin, John J. Gart, L. D. Meshalkin, A. R. Kagan, M. Zelen, R. E. Barlow, Jack Kalbfleisch, R. L. Prentice and Norman Breslow, and a reply by D. R. Cox.
- Damien, P. and Walker, S. (2002), “A Bayesian non-parametric comparison of two treatments,” *Scand. J. Statist.*, 29, 51–56.
- Doksum, K. (1974), “Tailfree and neutral random probabilities and their posterior distributions,” *Ann. Probability*, 2, 183–201.
- Dykstra, R. L. and Laud, P. (1981), “A Bayesian nonparametric approach to reliability,” *Ann. Statist.*, 9, 356–367.
- Elerian, O., Chib, S., and Shephard, N. (2001), “Likelihood inference for discretely observed nonlinear diffusions,” *Econometrica*, 69, 959–993.
- Ferguson, T. S. (1974), “Prior distributions on spaces of probability measures,” *Ann. Statist.*, 2, 615–629.
- Ferguson, T. S. and Phadia, E. G. (1979), “Bayesian nonparametric estimation based on censored data,” *Ann. Statist.*, 7, 163–186.
- Freireich, E. O. (1963), “The effect of 6 mercaptopurine on the duration of steroid induced remission in acute leukemia,” *Blood*, 21, 699–716.
- Gehan, E. A. (1965), “A generalized Wilcoxon test for comparing arbitrarily singly-censored samples,” *Biometrika*, 52, 203–223.
- Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1995), “Efficient parameterisations for normal linear mixed models,” *Biometrika*, 82, 479–488.

- (1996), “Efficient parametrizations for generalized linear mixed models,” in *Bayesian statistics, 5 (Alicante, 1994)*, New York: Oxford Univ. Press, Oxford Sci. Publ., pp. 165–180.
- Gjessing, H. K., Aalen, O. O., and Hjort, N. L. (2003), “Frailty models based on Lévy processes,” *Adv. in Appl. Probab.*, 35, 532–550.
- Hills, S. E. and Smith, A. F. M. (1992), “Parameterization issues in Bayesian inference,” in *Bayesian statistics, 4 (Peñíscola, 1991)*, New York: Oxford Univ. Press, pp. 227–246.
- Hjort, N. L. (1990), “Nonparametric Bayes estimators based on beta processes in models for life history data,” *Ann. Statist.*, 18, 1259–1294.
- Ishwaran, H. and James, L. F. (2004), “Computational methods for multiplicative intensity models using weighted gamma processes: proportional hazards, marked point processes, and panel count data,” *J. Amer. Statist. Assoc.*, 99, 175–190.
- Kalbfleisch, J. D. (1978), “Non-parametric Bayesian analysis of survival time data,” *J. Roy. Statist. Soc. Ser. B*, 40, 214–221.
- Kloeden, P. E. and Platen, E. (1992), *Numerical solution of stochastic differential equations*, vol. 23 of *Applications of Mathematics (New York)*, Berlin: Springer-Verlag.
- Laud, P. W., Damien, P., and Smith, A. F. M. (1998), “Bayesian nonparametric and covariate analysis of failure time data,” in *Practical nonparametric and semiparametric Bayesian statistics*, New York: Springer, vol. 133 of *Lecture Notes in Statist.*, pp. 213–225.
- Lo, A. Y. and Weng, C.-S. (1989), “On a class of Bayesian nonparametric estimates. II. Hazard rate estimates,” *Ann. Inst. Statist. Math.*, 41, 227–245.
- Muers, M. F., Shevlin, P., and Brown, J. (1996), “Prognosis in lung cancer: physicians’s opinions compared with outcome and a predictive model,” *Thorax*, 51, 894–902.
- Myers, L. E. (1981), “Survival functions induced by stochastic covariate processes,” *J. Appl. Probab.*, 18, 523–529.
- Papaspiliopoulos, O., Roberts, G. O., and Sköld, M. (2003), “Non-centered parameterizations for hierarchical models and data augmentation,” in *Bayesian statistics, 7 (Tenerife, 2002)*, New York: Oxford Univ. Press, pp. 307–326, with a discussion by Alan E. Gelfand, Ole F. Christensen and Darren J. Wilkinson, and a reply by the authors.
- (2007), “A general framework for the parametrization of Hierarchical models,” *Statist. Sci.*, 22, 59–73.
- R Development Core Team (2007), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
- Roberts, G. O. and Stramer, O. (2001), “On inference for partially observed nonlinear diffusion models using the Metropolis-Hastings algorithm,” *Biometrika*, 88, 603–621.
- Rogers, L. C. G. and Williams, D. (2000), *Diffusions, Markov processes, and martingales. Vol. 2*, Cambridge Mathematical Library, Cambridge: Cambridge University Press, itô calculus, Reprint of the second (1994) edition.
- Shephard, N. and Pitt, M. K. (1997), “Likelihood analysis of non-Gaussian measurement time series,” *Biometrika*, 84, 653–667.
- Stroock, D. W. and Varadhan, S. R. S. (2006), *Multidimensional diffusion processes*, Classics in Mathematics, Berlin: Springer-Verlag, reprint of the 1997 edition.
- Susarla, V. and Van Ryzin, J. (1976), “Nonparametric Bayesian estimation of survival curves from incomplete observations,” *J. Amer. Statist. Assoc.*, 71, 897–902.
- Wei, L. J. (1984), “Testing goodness of fit for proportional hazards model with censored observations,” *J. Amer. Statist. Assoc.*, 79, 649–652.
- Woodbury, M. A. and Manton, K. G. (1977), “A random-walk model of human mortality and aging,” *Theoret. Population Biology*, 11, 37–48.

- Xu, R. and O'Quigley, J. (2000), "Proportional hazards estimate of the conditional survival function," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 62, 667–680.
- Yashin, A. I. (1985), "Dynamics of survival analysis: conditional Gaussian property versus the Cameron-Martin formula," in *Statistics and control of stochastic processes (Moscow, 1984)*, New York: Optimization Software, Transl. Ser. Math. Engrg., pp. 466–485.
- Yashin, A. I. and Vaupel, J. W. (1986), "Measurement and estimation in heterogeneous populations," in *Immunology and epidemiology (Mogilany, 1985)*, Berlin: Springer, vol. 65 of *Lecture Notes in Biomath.*, pp. 198–206.