View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Research Papers in Economics

INTERNATIONAL CENTRE FOR ECONOMIC RESEARCH

ICER

WORKING PAPER SERIES

Pierpaolo De Blasi and Nils L. Hjort

THE BERNSTEIN-VON MISES THEOREM IN
SEMIPARAMETRIC COMPETING RISKS MODELS

Working Paper no. 17/2007
April 2007

APPLIED MATHEMATICS AND QUANTITATIVE METHODS
WORKING PAPER SERIES

The Bernstein-von Mises theorem in semiparametric competing risks models

Pierpaolo De Blasi* University of Turin

Nils L. Hjort University of Oslo

Abstract

Semiparametric Bayesian models are nowadays a popular tool in survival analysis. An important area of research concerns the investigation of frequentist properties of these models. In this paper, a Bernstein-von Mises theorem is derived for semiparametric Bayesian models of competing risks data. The cause-specific hazard is taken as the product of the conditional probability of a failure type and the overall hazard rate. We model the conditional probability as a smooth function of time and leave the cumulative overall hazard unspecified. A prior distribution is defined on the joint parameter space, which includes a beta process prior for the cumulative overall hazard. We show that the posterior distribution for any differentiable functional of interest is asymptotically equivalent to the sampling distribution derived from maximum likelihood estimation. A simulation study is provided to illustrate the coverage properties of credible intervals on cumulative incidence functions.

Keyword: Bayesian nonparametrics, Bernstein–von Mises theorem, beta process, competing risks, conditional probability of a failure type, semiparametric inference.

1 Introduction

The study of Bernstein-von Mises (BvM) type theorems has recently received a renewed interest in the context of nonparametric statistics. Indeed, such an interest stems from the debate on posterior inconsistency originated by the paper of Diaconis and Freedman (1986). The BvM theorem states that the posterior distribution of the model parameter centered at the maximum likelihood estimator (MLE) is asymptotically equivalent to the sampling distribution of the MLE. In a parametric setup, it represents a quite standard result, roughly implied by the consistency and asymptotic normality of the

*Address for correspondence: pierpaolo.deblasi@unito.it

MLE. See, e.g., Schervish (1995) and references therein. On the other hand, in infinite-dimensional models, the choice of the prior distribution influences the large sample properties of the posterior and BvM-type results are in general difficult to establish, because of both the over-whelming mathematics involved in their derivation and the fact they may not hold.

Cox (1993) provides an example of nonparametric regression with a Gaussian prior where posterior inference and maximum likelihood inference disagree asymptotically. His arguments are further developed by Freedman (1999). Somehow surprisingly, in infinite-dimensional models the BvM theorem is not even guaranteed by the consistency of the posterior, since problems may arise due to a suboptimal convergence rate. See Zhao (2000). For nonparametric survival models there exist positive results for the family of neutral to the right processes (Doksum 1974), which constitute the most common prior used on the space of survival distributions. Indeed, Kim and Lee (2004) investigate the BvM theorem for right-censored survival data and show that it holds under minimal conditions that are matched by the most common processes within this family. Their method is based on describing neutral to the right processes via the corresponding random cumulative hazard, taken as an *increasing additive process*, i.e. an increasing process with independent and not necessarily stationary increments. See, e.g., Sato (1999). This approach was first introduced by Hjort (1990) and developed further by Kim (1999) in terms of counting processes. Kim and Lee (2004) proceed by proving that the BvM theorem holds for the cumulative hazard function and then extend the result to the survival distribution via the functional delta method. The survival function is in fact recovered from the cumulative hazard via the product integration operator, which is compactly differentiable. The asymptotic normality of the survival function is then expressed in terms of the sampling distribution of the Kaplan-Meier estimator. Still in the context of nonparametric survival models, the following step in the analysis of BvM type asymptotics is naturally represented by the proportional hazards regression model, which stands out for its wide use in applications. The derivation of the BvM theorem within this framework has been successfully faced in Kim (2006) and De Blasi and Hjort (2007). In fact, the convenience of working at the cumulative hazard level is that Bayesian methods can be readily extended to more complex event history data by using the multiplicative intensity model of Aalen (1978). For a comprehensive treatment of Aalen's approach, see the monograph by Andersen, Borgan, Gill and Keiding (1993) (ABGK henceforth).

In this paper we derive a BvM result for competing risks models, which represent another important class of statistical tools in the context of event history analysis, aiming at the description of the occurrences of failure times with multiple endpoints. See Lawless (2003, Chapter 9) for an exhaustive account on competing risks. Here we consider a particular semiparametric formulation and study the asymptotic normality of posterior distributions for quantities derived from the model parameters via differentiable functionals. This formulation has statistical interest in its own and is described below.

We consider the pair (T^0, D^0), where T^0 is the failure time and $D^0 \in \{1, \ldots, k\}$ is the type of failure. This means that the event under observation has k different and

mutually exclusive outcomes. The joint distribution of (T^0, D^0) can be specified via the cause-specific hazard (CSH) function:

$$\alpha_j(t) = \lim_{\Delta t \to 0} \Pr\{t \le T^0 < t + \Delta t,\ D^0 = j | T \ge t\}/\Delta t, \quad j = 1, \ldots, k,$$

that is $\alpha_j(t)$ is the instantaneous rates of a failure of type j at time t. Alternatively, one can describe competing risk data via the marginal probabilities $P_j(t) = \Pr\{T^0 \le t, D^0 = j\}$ also known as cumulative incidence function (CIF). We rather consider cumulative CSH, defined as $\mathrm{d}A_j(t) = \mathrm{d}P_j(t)/S(t-)$, where $S(t) = \Pr\{T^0 > t\}$ is the survival function. In the continuous case, A_j is the integral of α_j, whereas in general it is a right-continuous, non decreasing function with jumps in $[0, 1]$. We work in a semiparametric setting and specify A_j as follows:

$$A_j(t) = \int_0^t p_j(s, \theta) \mathrm{d}A(s), \quad j = 1, \ldots, k, \tag{1.1}$$

where (i) $p_j : \mathbb{R}_+ \times \mathbb{R}^p \to [0, 1]$ is a smooth function continuous in t and twice differentiable in θ such that $\sum_{j \le k} p_j(s, \theta) = 1$ for each s and θ; (ii) $A = \sum_{j \le k} A_j$ is the cumulative overall hazard and is left unspecified. Our results apply to a general class of models, for p_j as detailed in the following equation (7.1). However, in order to ease the flow of ideas and avoid heavy notation, we focus first on the crucial case

$$\begin{cases} p_j(s, \theta) = & e^{\theta_{j1} + \theta_{j2} s} p_k(s, \theta), \qquad j = 1, \ldots, k-1 \\ p_k(s, \theta) = & \left\{ \sum_{h \le k} e^{\theta_{h1} + \theta_{h2} s} \right\}^{-1} \end{cases} \tag{1.2}$$

where $\theta_{k1} = 0, \theta_{k2} = 0$ is needed for guaranteeing the identifiability of the model. The proportionality factor $p_j(t, \theta)$ describes the conditional probability of a failure of type j at time t, given one failure occurs at that time:

$$p_j(t, \theta) = \lim_{\Delta t \downarrow 0} \frac{\Pr\{T^0 \in [t, t + \Delta t),\ D^0 = j\}}{\Pr\{T^0 \in [t, t + \Delta t)\}} = \frac{\mathrm{d}A_j(t)}{\mathrm{d}A(t)}, \tag{1.3}$$

see ABGK Section II.6. Following the terminology of Gasbarra and Karia (2000), we call p_j the *cause-specific conditional probability*. Note that it corresponds to the subdensity $\mathrm{d}P_j/\mathrm{d}t$ normalized over its sum. In particular, (1.2) can be seen as the cause-specific conditional probabilities that arise by starting with Gompertz-type CSH and by normalization for the kth one. In order to avoid confusion, the reader should note the difference with the conditional probability $\Pr(T^0 \le t \,|\, D^0 = j)$, sometimes used for describing competing risks data in the so called mixture model, see e.g. Larson and Dinse (1985).

The proposed method consists in carrying out Bayesian estimation of the joint distribution of (T^0, D^0) via the cause-specific conditional probability p_j and the cumulative overall hazard A. This entails the specification of a prior density function for θ and a prior distribution for the functional parameter A. The latter is accomplished by resorting to the beta process of Hjort (1990), an increasing additive process widely used for modeling cumulative hazards. A related approach can be found in Gasbarra

and Karia (2000), who also consider the cause-specific conditional probability and the overall hazard rate, but with a different prior specification. As for the cause-specific conditional probability, they take a random partition of the time axis and assign on each segment an independent k-dimensional Dirichlet distributed random vector. Then p_j is obtained by kernel smoothing. As for the overall hazard, they propose to use a convolution of gamma processes at the hazard rate level, in the spirit of Lo and Weng (1989). Another approach to competing risks in a Bayesian nonparametric framework is the one of Salinas-Torres et al. (2002), who model directly the CIF's via a Dirichlet multivariate process prior (see definition therein). Their approach is different in that they assume a system of independent latent failure times and exploit the representation of the CIF as the product $\Pr(T^0 \leq t \,|\, D^0 = j) \times \Pr(D^0 = j)$.

Our model is semiparametric, with (A, θ) taking value in the product space $\Omega = D_{[0,\tau]} \times \mathbb{R}^p$, where $p = 2k - 2$ and $D_{[0,\tau]}$ is the space of cadlag function on the (possibly infinite) time interval $[0, \tau]$. All the relevant survival quantities are obtained from (A, θ) via functionals like the cumulative CSH A_j (see (1.1)), the survival function $S(t) = \prod_{[0,t]}\{1 - \mathrm{d}A(s)\}$ and the CIF $P_j(t)$, in the present setting given by

$$P_j(t) = \int_0^t \prod_{[0,s)} \{1 - \mathrm{d}A(u)\} p_j(s, \theta) \mathrm{d}A(s), \quad j = 1, \ldots, k. \tag{1.4}$$

Here $\prod$ stands for the product integral, see Gill and Johansen (1990). As we are interested in the asymptotic properties of the posterior distribution of A_j, S and P_j, we make use of the following two main facts: (i) for A_j given in (1.1), the mapping $\Psi : (A, \theta) \to (A_1, \ldots, A_k)$ from Ω to $(D_{[0,\tau]})^k$ is compactly differentiable, (ii) compact differentiability satisfies the chain rule: the composition of differentiable functionals is differentiable, with derivative equal to the composition of the derivatives (see Gill 1989). It follows that S and P_j are compactly differentiable functionals of (A, θ). A noteworthy consequence is that the BvM theorems for the posterior distributions of A_j, S and P_j are implied by the asymptotic normality of the posterior distribution of (A, θ) under appropriate priors. Hence, it suffices to prove the BvM for (A, θ). The formulation in (1.1) is convenient in that, as it will be shown in Section 2, the total likelihood factorizes into two parts and estimation of A and θ can be carried out separately. Moreover, this alleviates remarkably the difficulties usually encountered in describing the full posterior in semiparametric inference: with independent priors for A and θ we get independence also with respect to their posterior distributions.

The paper is organized as follows. In Section 2, we introduce the counting process formulation of cause-specific failure times and exploit the similarity of $p_j(t, \theta)$ in (1.2) with a multinomial logistic regression model in order to obtain the asymptotic properties of the likelihood estimators of A and θ. Then, the allied limiting distributions for A_j, S and P_j are derived using the functional delta method. In Section 3, we develop the Bayesian treatment of the model by describing the prior distribution for (A, θ), by deriving the form of the posterior distribution together with sampling schemes for posterior averaging. Section 4 provides the main result of the paper, namely the BvM theorem for the posterior distribution of (A, θ). In Section 5, a simulation study is presented, with focus on the empirical coverage of credible sets on CIF's. In order to

ease the flow of ideas, we collect proofs and technical lemmas in Section 6. In Section 7 we provide conditions for the BvM theorems to hold for a large family of models for p_j, which include (1.2) as special case; we also show how the previously developed techniques may be adapted to this framework. Finally, some concluding remarks and lines of future research are provided.

2 Asymptotics for likelihood estimation

Let us start by introducing the counting process formulation of competing risks data. As we want to account for right-censoring, we indicate the sample by $(T_1, D_1), \ldots, (T_n, D_n)$, where T_i is the (possibly right censored) failure time and $D_i = \{0, 1, \ldots, k\}$ is the observed failure type: $D_i = 0$ if T_i is right censored, whereas $D_i = j$ if i-th individual is observed to fail due to cause j. The censoring mechanism is assumed to be independent of the failure time and the failure type distribution. For each observation (T_i, D_i), we consider k counting processes $N_{i,j}(t) = I\{T_i \leq t, D_i = j\}$, and the at-risk process $Y_i(s) = I\{T_i \geq t\}$. According to Aalen's multiplicative intensity model, $N_{i,j}$ can be decomposed as

$$N_{i,j}(t) = \int_0^t Y_i(s)\mathrm{d}A_j(s) + M_{i,j}(t), \quad j = 1, \ldots, k, \tag{2.1}$$

i.e. the sum of the intensity process and a martingale residual $M_{i,j}$. In the following we use the notation $\langle M \rangle$ for the variation process of M and $\langle M, M' \rangle$ for the covariation process of M and M', M and M' being martingales. Under standard regularity conditions, that we assume to hold, we have $\langle M_{i,j} \rangle(t) = \int_0^t Y_i(s)\mathrm{d}A_j(s)$ and orthogonality at the failure type level: $\langle M_{i,j}, M_{i,h} \rangle = 0$. Summation at the individual level preserves the representation in (2.1), as well as orthogonality, and will be denoted by suppressing the index i. We also make use of the aggregated process $N_{\cdot}(t) = \sum_{j \leq k} N_j(t)$, which counts the number of failure of any types occurred before time t.

The likelihood for (A, θ) can be expressed by the product integral in a multinomial form:

$$L(A, \theta) = \prod_t \left\{ \prod_{j \leq k} (Y(t)\mathrm{d}A_j(t))^{\mathrm{d}N_j(t)} \Big\{ 1 - \sum\nolimits_{j \leq k} \mathrm{d}A_j(t) \Big\}^{1 - \mathrm{d}N_{\cdot}(t)} \right\}$$

cf. equation (2.7.2") in ABGK, Section II.7. Upon substitution of (1.1), the likelihood for θ can be separated out from that of A, so that the $L(A, \theta)$ factorizes in two components, say $L(A)$ and $L(\theta)$, where

$$L(A) = \prod_t \Big\{ \mathrm{d}A(t)^{\mathrm{d}N_{\cdot}(t)} \big\{ 1 - \mathrm{d}A(t) \big\}^{Y(t) - \mathrm{d}N_{\cdot}(t)} \Big\}, \tag{2.2}$$

$$L(\theta) = \prod_{i \leq n} \prod_{j \leq k} p_j(\theta, t_i)^{\Delta N_{i,j}(t_i)}. \tag{2.3}$$

$L(A)$ is the likelihood of the cumulative overall hazard, and simply corresponds to the case of right-censored survival times, while $L(\theta)$ is the conditional likelihood for θ and

depends only on uncensored observations. The factorization of $L(A,\theta)$ represents one of the main property of model (1.1): it leads to frequentist estimation in a straightforward way by means of the Nelson-Aalen estimator $\widehat{A}_n(t) = \int_0^t Y(s)^{-1}\mathrm{d}N_\cdot(s)$ for A and of the MLE $\widehat{\theta}_n$ for θ. Upon substitution of (1.2) for $p_j(\theta,t)$ in (2.3), the conditional likelihood $L(\theta)$ resembles the likelihood for a multinomial logistic regression model with the type of failure as response variable and the time of failure as regression variable. Then, one can exploit a well known result of the multinomial logistic regression model, namely that the likelihood of the regression coefficient is strictly concave if there is overlap on the space of covariate variables, see Albert and Anderson (1984). As we will show in the sequel, this allows in our setting to provide weak and neat conditions for the large sample properties of likelihood estimation to hold.

To this aim, we postulate that model (1.1)-(1.2) is in force for a certain overall hazard rate $\alpha(\cdot) = \alpha_{\mathrm{tr}}(\cdot)$, with cumulative hazard $A_{\mathrm{tr}}(t) = \int_0^t \alpha_{\mathrm{tr}}(s)\mathrm{d}s$, and for a parameter vector $\theta_{\mathrm{tr}} \in \mathbb{R}^{2k-2}$. We also assume that observations are recorded over a fixed and finite time window $[0,\tau]$. The conditions required are the following:

(A) There exists a positive $y(\cdot)$ such that $\sup_{t\in[0,\tau]} |n^{-1}Y(t) - y(t)| \to_p 0$.

(B) $\int_0^\tau (t^2+1)^3 \alpha_{\mathrm{tr}}(t)\mathrm{d}t < \infty$.

(C) $\exists j,h \in \{1,\ldots,k\}$, $j \neq h$, such that $p_j(t,\theta_{\mathrm{tr}}), p_h(t,\theta_{\mathrm{tr}}) > 0$ for any t.

Condition (A) guarantees that $Y(\tau) \to \infty$ in probability as $n \to \infty$, which is needed for the asymptotic distribution of the Nelson-Aalen estimator $\widehat{A}_n$. For example, if right censoring is determined by independent censoring times with common distribution function G, then it is sufficient to assume that $G(\tau-) < 1$. The integrability condition (B) is needed for the convergence (in probability) of the variation processes of martingales obtained from M_j via stochastic integration. Condition (C) assures that, for sufficiently large sample sizes, there is overlap of cause-specific failure times, which is needed for the concavity of $L(\theta)$, see Lemma 6.1 in Section 6.

Before describing the limiting distribution of $(\widehat{A}_n, \widehat{\theta}_n)$, we introduce some more notation that will be also needed in the rest of the paper. Let $\ell_n(\theta) = \log L(\theta)$ and use the counting process formulation to write

$$\ell_n(\theta) = \sum_{j\leq k} \int_0^\tau \big[z_j(t)^{\mathrm{t}}\theta + \log p_k(t,\theta)\big]\mathrm{d}N_j(t), \tag{2.4}$$

where $z_j(t)$ is the $(2k-2)$-dimensional function having $(1,t)^{\mathrm{t}}$ in the j-th block and zeros elsewhere. Note that $z_k(t)$ is identically zero and is introduced for mathematical convenience. Upon definition of $e(t,\theta) = \sum_{j\leq k} z_j(t)p_j(t,\theta)$, the function $z_j(t) - e(t,\theta)$ stands for $\partial \log p_j(t,\theta)/\partial\theta$. Next, define the information matrix $\Sigma = \int_0^\tau V(t,\theta_{\mathrm{tr}})y(t)\alpha_{\mathrm{tr}}(t)\mathrm{d}t$, where $V(t,\theta) = \sum_{j\leq k} p_j(t,\theta)[z_j(t) - e(t,\theta)]^{\otimes 2}$, so that $\upsilon \sim \mathrm{N}(0,\Sigma^{-1})$ denotes a multivariate normal random vector with covariance matrix given by the inverse of Σ. As for the asymptotic distribution of the Nelson-Aalen estimator $\widehat{A}_n$, define the function $\sigma^2(t) = \int_0^t y(s)^{-1}\alpha_{\mathrm{tr}}(s)\mathrm{d}s$ and let W be a standard Brownian motion. For a vector

$b = (b_1, \ldots, b_p)$, write $|b| = (b^{\mathrm{t}} b)^{1/2}$ and $b^{\otimes 2} = b\, b^{\mathrm{t}}$, while, for B a $p \times p$ matrix, $|B|$ is the determinant and $\mathrm{vec}(B)$ is the column vector of length p^2 with the jth block equal to the jth column of B. The covariance matrix of a matrix-valued random variable $\mathbf{X}$ is then defined as $\mathrm{cov}(\mathbf{X}) = \mathrm{E}\big\{[\mathrm{vec}(\mathbf{X}) - \mathrm{vec}(\mathrm{E}\mathbf{X})][\mathrm{vec}(\mathbf{X}) - \mathrm{vec}(\mathrm{E}\mathbf{X})]^{\mathrm{t}}\big\}$. Finally, let $C_{[0,\tau]}$ be the space of continuous functions defined on the interval $[0, \tau]$.

THEOREM 2.1 *Assume conditions (A)-(C) hold. Then* $\sqrt{n}(\widehat{\theta}_n - \theta_{\mathrm{tr}})$ *and* $\sqrt{n}(\widehat{A}_n - A_{\mathrm{tr}})$ *are asymptotically independent and*

$$\sqrt{n}(\widehat{A}_n - A_{\mathrm{tr}}) \quad \to_d \quad \mathrm{W}(\sigma^2) \quad \text{on } D_{[0,\tau]}. \tag{2.5}$$

$$\sqrt{n}(\widehat{\theta}_n - \theta_{\mathrm{tr}}) \quad \to_d \quad \mathrm{N}(0, \Sigma^{-1}) \tag{2.6}$$

The asymptotic independence of $\sqrt{n}(\widehat{\theta}_n - \theta_{\mathrm{tr}})$ and $\sqrt{n}(\widehat{A}_n - A_{\mathrm{tr}})$ is a key ingredient for the derivation of the asymptotic distribution of the plug-in estimators for A_j, S and P_j via the functional delta method. As for A_j, consider the mapping $\Psi : (A, \theta) \to (A_1, \ldots, A_k)$ from $D_{[0,\tau]} \times \mathbb{R}^{2k-2}$ to $(D_{[0,\tau]})^k$ defined by equations (1.1) and (1.2) and write Ψ as the composition $\Psi = \Psi_2 \circ \Psi_1$, where $\Psi_1 : \mathbb{R}^{2k-2} \to (C_{[0,\tau]})^k$ is defined by $\Psi_1(\theta) = \big[p_1(t,\theta), \ldots, p_k(t,\theta)\big]^{\mathrm{t}}$ and is compactly differentiable at each point of $x \in \mathbb{R}^{2k-2}$ with derivative given by

$$(\mathrm{d}\Psi_1(\theta) \cdot x)(t) = \begin{pmatrix} [z_1(t) - e(t,\theta)]^{\mathrm{t}} x p_1(t,\theta) \\ \vdots \\ [z_k(t) - e(t,\theta)]^{\mathrm{t}} x p_k(t,\theta) \end{pmatrix}$$

On the other hand, $\Psi_2 : (C_{[0,\tau]})^k \times D_{[0,\tau]} \to (D_{[0,\tau]})^k$ is defined by $\Psi_2(p_1, \ldots, p_k, A) = (A_1, \ldots, A_k)$, where $A_j = \int p_j \mathrm{d}A$. Using the properties of the integration operator (see ABGK Proposition II.8.6) we have that, for $(h_1, \ldots, h_k, H)$ in the space $(C_{[0,\tau]})^k \times D_{[0,\tau]}$ such that each h_j is integrable with respect to H, Ψ_2 has derivative:

$$\mathrm{d}\Psi_2(p_1, \ldots, p_k, A) \cdot (h_1, \ldots, h_k, H) = \begin{pmatrix} \int h_1 \mathrm{d}A + \int p_1 \mathrm{d}H \\ \vdots \\ \int h_k \mathrm{d}A + \int p_k \mathrm{d}H \end{pmatrix}$$

The compact differentiability of Ψ is then a direct consequence of the chain rule of compact differentiability. The next corollary gives the limiting distribution of $\sqrt{n}[\widehat{A}_j(t) - A_j(t)]$, $j = 1, \ldots, k$, in terms of a vector of dependent Gaussian processes. Here $\widehat{A}_j$ stays for A_j in (1.1) with $(\widehat{A}_n, \widehat{\theta}_n)$ substituted for (A, θ). The thesis follows from two applications of the functional delta method, see Gill (1989).

COROLLARY 2.1 *For* $j = 1, \ldots, k$ *define the Gaussian process*

$$U_j(t) = \int_0^t [z_j(s) - e(s, \theta_{\mathrm{tr}})]^{\mathrm{t}} \upsilon\, p_j(s, \theta_{\mathrm{tr}}) \alpha_{\mathrm{tr}}(s) \mathrm{d}s + \int_0^t p_j(s, \theta_{\mathrm{tr}}) \mathrm{d}W\big(\sigma^2(s)\big).$$

Then, the following weak convergence result on $(D_{[0,\tau]})^k$ *holds:*

$$\sqrt{n}\big[\ (\widehat{A}_1, \ldots, \widehat{A}_k) - (A_1, \ldots, A_k)\big] \to_d (U_1, \ldots, U_k).$$

In order to derive the asymptotic distribution of S and P_j, it is convenient to look at competing risks data as the transition times of a Markov process with one transient state "0 : alive" and absorbing state $j = 1, \ldots, k$ corresponding to "failure of type j". Let $\mathbf{P}(s,t)$ be the corresponding transition matrix for $0 \leq s \leq t$ and write $\mathbf{P}(t) = \mathbf{P}(0,t)$. Then

$$\mathbf{P}(t) = \prod_{[0,t]} \{\mathbf{I} + \mathrm{d}\mathbf{A}(s)\} \tag{2.7}$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{A}$ is the transition intensity matrix with element $(1,1)$ equal to $-A$, elements $(1, j+1)$ equal to A_j, elements $(j+1, j+1)$ equal to 1 and zeros elsewhere, j running trough $\{1, \ldots, k\}$. Note that $[S(t), P_1(t), \ldots, P_k(t)]$ corresponds to the first row of $\mathbf{P}(t)$. The thesis of Corollary 2.1 can be written as $\sqrt{n}(\widehat{\mathbf{A}} - \mathbf{A}) \to_d \mathbf{U}$ for $\mathbf{U}$ the matrix-valued Gaussian process with first row equal to $(-W(\sigma^2), U_1, \ldots, U_n)$ and zeros elsewhere. Upon definition of the matrix C_j having element $(1,1)$ equal to -1, element $(1, j+1)$ equal to 1 and zeros elsewhere, the covariance matrix of $\mathbf{U}(t)$ is given by $\mathrm{cov}(\mathbf{U}(t)) = \sum_{j,h \leq k} \mathrm{vec}(C_j)\mathrm{vec}(C_h)^{\mathrm{t}} \omega_{jh}(t)$, where

$$\begin{aligned}\omega_{jh}(t) = &\int_0^t [z_j(s) - e(s,\theta_{\mathrm{tr}})]^{\mathrm{t}} p_j(s,\theta_{\mathrm{tr}}) \alpha_{\mathrm{tr}}(s) \mathrm{d}s \, \Sigma^{-1} \int_0^t [z_h(s) - e(s,\theta_{\mathrm{tr}})] p_h(s,\theta_{\mathrm{tr}}) \alpha_{\mathrm{tr}}(s) \mathrm{d}s \\ &+ \int_0^t p_j(s,\theta_{\mathrm{tr}}) p_h(s,\theta_{\mathrm{tr}}) \frac{\alpha_{\mathrm{tr}}(s)}{y(s)} \, \mathrm{d}s\end{aligned}$$

We now let the previous notation prevail. The next corollary describes the limiting distribution of $\sqrt{n}[\widehat{\mathbf{P}}(t) - \mathbf{P}(t)]$, where $\widehat{\mathbf{P}}$ stays for the transition matrix $\mathbf{P}$ with $(\widehat{A}_n, \widehat{\theta}_n)$ substituted in $\mathbf{A}$ in (2.7) via equation (1.1). The thesis follows from an application of the functional delta method and the formula for the derivative of the product integral, see Gill and Johansen (1990).

COROLLARY 2.2 *The transition matrix $\widehat{\mathbf{P}}$ has the following asymptotic behavior:*

$$\sqrt{n}[\widehat{\mathbf{P}}(t) - \mathbf{P}(t)] \to_d \mathbf{Z}(t) = \int_0^t \mathbf{P}(s-) \, \mathrm{d}\mathbf{U}(s) \, \mathbf{P}(s,t)$$

for each $t \in [0, \tau]$, where $\mathbf{Z}(t)$ has covariance matrix given by

$$\mathrm{cov}(\mathbf{Z}(t)) = \int_0^t \mathbf{P}(s,t)^{\mathrm{t}} \otimes \mathbf{P}(s-) \, \mathrm{d}\, \mathrm{cov}(\mathbf{U}(s)) \, \mathbf{P}(s,t) \otimes \mathbf{P}(s-)^{\mathrm{t}}.$$

It follows that an explicit formula for the asymptotic distribution of $\sqrt{n}[\widehat{P}_j(t) - P_j(t)]$ is as follows:

$$Z_j(t) = \int_0^t S(u-) \mathrm{d}U_j(u) - \int_0^t \frac{P_j(t) - P_j(u)}{S(u)} S(u-) \mathrm{dW}(\sigma^2(u)),$$

with asymptotic variance

$$\begin{aligned}\mathrm{var}(Z_j(t)) \;=\; &\int_0^t S(u-)^2 \Big\{ \frac{\big(P_j(t) - P_j(u)\big)^2}{S(u)^2} \, \sigma^2(u) \mathrm{d}u \\ &-2 \frac{P_j(t) - P_j(u)}{S(u)} p_j(u,\theta_{\mathrm{tr}}) \sigma^2(u) \mathrm{d}u + \mathrm{d}\omega_{jj}(u) \Big\}.\end{aligned} \tag{2.8}$$

3 Bayesian inference

As far as model (1.1) is concerned, Bayesian inference requires the specification of a prior distribution for (A, θ) on the infinite-dimensional space $D_{[0,\tau]} \times \mathbb{R}^{2k-2}$. We proceed by taking A and θ to be independent and, then, derive the posterior distributions separately because of the factorization in the total likelihood $L(A, \theta)$.

As for the prior distribution of A, we resort to the beta process of Hjort (1990), which is an increasing additive process whose paths lie, with probability one, in the space of cumulative hazard functions. It is worth noting that a fully nonparametric treatment of the competing risks model was implicit in Hjort (1990, Section 5), who showed how to use the beta process for Bayesian inference on nonhomogeneous Markov processes. His proposal was to assign independent beta processes to all cumulative transition intensities, using the fact that, given a sample of right-censored transition times, they preserve independence and are still beta distributed in the posterior. A formal definition of the beta process is as follows. Let $\mathcal{A}$ is the set of right-continuous non-decreasing functions A on $\mathbb{R}_+$, such that $A(0) = 0$ and A is increasing to infinity with jumps in $[0, 1]$. Let $A_0 \in \mathcal{A}$ with jumps at points $\{t_1, t_2, \ldots\}$ and let $c(\cdot)$ be a piecewise continuous, nonnegative real valued function on $[0, \infty)$. Then, a beta process with parameters c and A_0, in symbols $A \sim \text{Beta}(c, A_0)$, is defined as an increasing additive process with Lévy-Khinchine representation given by

$$\mathrm{E}\big[e^{-uA(t)}\big] = \prod_{j:\, t_j \le t} \mathrm{E}\big[e^{-u\Delta A(t_j)}\big] e^{-\int_0^t \int_0^1 (1-e^{-us}) s^{-1} (1-s)^{c(z)-1} c(z) \mathrm{d}s \, \mathrm{d}A_{0,c}(z)},$$

where $\Delta A(t_j)$ is the jump size at location t_j and is distributed as a beta random variable of parameters $c(t_j)\Delta A_0(t_j)$ and $c(t_j)\{1 - \Delta A_0(t_j)\}$, whereas $A_{0,c}(t) = A_0(t) - \sum_{t_j \le t} \Delta A_0(t_j)$. Note that $\mathrm{d}A(s)$ has mean $\mathrm{d}A_0(s)$ and variance $\mathrm{d}A_0(s)\{1-\mathrm{d}A_0(s)\}/\{1+c(s)\}$, indicating that A_0 is the prior guess at A and c determines the concentration of the random function around A_0. In particular, the choice $c(t) = m \exp\{-A_0(t)\}$ makes the random distribution function $1-\pi_{[0,t]}\{1-\mathrm{d}A(s)\}$ distributed as a Dirichlet process with prior mean equal to $\big[1 - \pi_{[0,t]}\{1 - \mathrm{d}A_0(s)\}\big]$ and concentration parameter m. In this case m can be interpreted as the strength of the prior beliefs in A_0, corresponding to the size of an imaginary prior sample from a lifetime distribution with cumulative hazard A_0.

The convenience of using a beta process prior for the overall cumulative hazard consists in the fact that we are modeling the occurrence of simple events, disregarding the type of failure observed. In fact, the likelihood contribution $L(A)$ depends only on the failure times $t_1, \ldots, t_n$ and the censoring mechanism (see equation (2.2)). Then, one can exploit the conjugacy of the beta process for i.i.d. right-censored data to get

$$A \,|\, \text{data} \sim \text{Beta}\left\{c + Y, \int \frac{c\,\mathrm{d}A_0 + \mathrm{d}N_\cdot}{c + Y}\right\}, \tag{3.1}$$

where we have used the counting process notation introduced in Section 2. The second parameter corresponds to the posterior mean and has increments given by a convex linear combination of the Nelson-Aalen estimator $\widehat{A}_n(t)$ and the prior mean A_0. The

plain Nelson-Aalen estimator $\widehat{A}_n$ arises in the noninformative case of $c(t) \equiv 0$ for any t. If A_0 is continuous, then, in the posterior, there are new fixed points of discontinuity at any observed failure time t_i, the distribution of the jump size being

$$\Delta A(t_i) \,|\, \text{data} \sim \text{beta}\Big(c(t_i)\triangle A_0(t_i) + \mathrm{d}N_\cdot(t_i), c(t_i)[1 - \triangle A_0(t_i)] + Y(t_i) - \mathrm{d}N_\cdot(t_i)\Big).$$

Sample paths from the posterior distribution of A are readily obtained by simulating independently the beta-distributed jumps and the continuous paths. For a fine grid of the time axis, the increments of the continuous part can be approximated by summing beta random variates, generated in accordance with (3.1).

As for the parameter θ, any density function $\pi(\theta)$ with support $\mathbb{R}^{2k-2}$ can be employed, although specifying a meaningful distribution can be a difficult task. A possibility is to adopt the prior density dictated by Jeffreys's rule, which is suited to the case of little available prior information. For general parametric models, the Jeffreys's prior is proportional to $|J_n(\theta)|^{1/2}$, the square root of the determinant of the information matrix. Since the conditional likelihood $L(\theta)$ is interpretable in terms of a multinomial logistic regression model, it is easy to see that Jeffreys's prior is given by

$$\pi(\theta) \propto \Big| \sum_{i:\, d_i \neq 0} V(t_i, \theta) \Big|^{1/2}. \tag{3.2}$$

See also Ibrahim and Laud (1991). In order to simulate from the posterior density $\pi(\theta|\text{data}) \propto \pi(\theta) \times L(\theta)$, we exploit the concavity of $L(\theta)$. In fact, by Lemma 6.1 in the Section 6, $L(\theta)$ is concave as long as there is overlapping of cause-specific failure times. In this case, a concave prior $\pi(\theta)$ leads to a concave posterior density, so that one can use coordinatewise the adaptive rejection sampling (Gilks and Wild, 1992).

Once we simulate A^* from $A \,|\, \text{data}$ and θ^* from $\pi(\theta|\text{data})$, inference on CIF's can be obtained via posterior averaging. The trajectory $\{A^*(t),\ t \geq 0\}$ will be of pure jumps, with jump times $0 < s_1 < s_2 < \ldots$ obtained by reordering the uncensored times of failure in the data and the time points used in the discretization of the continuous part. Then,

$$P_j^*(t) = \sum_{i:\ s_i \leq t} \Big\{ \prod_{l<i} [1 - \Delta A^*(s_l)] p_j(s_i, \theta^*) \Delta A^*(s_i) \Big\}.$$

is a variate from the posterior distribution of P_j, see equation (1.4).

4 Bernstein-von Mises theorem

In this section we investigate the BvM theorem for the posterior distribution of A_j and P_j. Indeed, we prove that, under the choice of beta process for the overall hazard and minimal conditions on the prior density $\pi(\theta)$, the posterior distribution of (A, θ) is asymptotically equivalent to the sampling distribution of the likelihood estimates $(\widehat{A}_n, \widehat{\theta}_n)$, as stated in Theorem 2.1. The claimed BvM result for A_j and P_j is, then, direct consequence of the functional delta method, as implemented in Corollary 2.1 and

2.2. It is worth noting that the same is true for all differentiable functionals of (A, θ) or of the CSH's A_j: we can then rely on posterior averaging having valid frequentist properties. Indeed, thanks to the BvM result, Bayesian credible sets are guaranteed to reach asymptotically nominal coverage probability like consistent estimation based on the likelihood. This has important consequences from a practical point of view. Since the Bayesian computational capacity has increased, Bayesian credible sets represent an alternative to confidence intervals when traditional methods do not lead to easily implementable algorithms. The BvM theorem is the theoretical justification of this practice.

As in Section 2, we assume that data are generated under model (1.1) with true parameters $(A_{\mathrm{tr}}, \theta_{\mathrm{tr}})$, and that conditions (A)-(C) hold. The convergence of the posterior distribution holds in probability, where convergence in probability refers to repeated sampling from the true distribution of (T^0, D^0).

THEOREM 4.1 *Let A be a beta process (A_0, c) and let θ, independent of A, have density π. Assume that A_0 is continuous with bounded and positive density on $(0, \tau)$ and that $0 < \inf_{t\in[0,\tau]} c(t) \leq \sup_{t\in[0,\tau]} c(t) < \infty$. Moreover, π is assumed to be positive and continuous at θ_{tr}. Then, as $n \to \infty$,*

$$\sqrt{n}(A - \widehat{A}_n) \,|\, \text{data} \quad \to_d \quad \mathrm{W}(\sigma^2) \quad \text{on } D_{[0,\tau]} \text{ in probability,} \tag{4.1}$$

$$\sqrt{n}(\theta - \widehat{\theta}_n) \,|\, \text{data} \quad \to_d \quad \mathrm{N}(0, \Sigma^{-1}) \text{ in probability.} \tag{4.2}$$

COROLLARY 4.1 *Under the hypothesis of Theorem 4.1*

$$\sqrt{n}\big[\ (A_1, \ldots, A_k) - (\widehat{A}_1, \ldots, \widehat{A}_k)\big] \,|\, \text{data} \to_d (U_1, \ldots, U_k)^{\mathrm{t}}$$

on $(D_{[0,\tau]})^k$ in probability, where $\widehat{A}_j(t)$ is defined as in Corollary 2.1.

COROLLARY 4.2 *Under the hypothesis of Theorem 4.1, for $j = 1, \ldots, k$,*

$$\sqrt{n}(P_j - \widehat{P}_j) \,|\, \text{data} \to_d \int_0^t S(u-)\mathrm{d}U_j(u) - \int_0^t \frac{P_j(t) - P_j(u)}{S(u)} S(u-)\mathrm{dW}(\sigma^2(u))$$

on $D_{[0,\tau]}$ in probability, where $\widehat{P}_j(t) = \int_0^t \pi_{[0,s)} \left\{1 - \mathrm{d}\widehat{A}_n(u)\right\} p_j(s, \widehat{\theta}_n) \mathrm{d}\widehat{A}_n(s)$.

5 Simulation study

In this section we investigate the validity of the BvM theorem on simulated data by comparing the coverage probability of likelihood and Bayesian intervals for the CIF. We generate competing risks data with two types of failure ($k = 2$) from independent lifetimes (T_1, T_2) having different Gompertz distributions. Specifically, $T_1 \sim \mathrm{Gomp}(a_1, b_1)$ with $(a_1, b_1) = (\log(0.05), 0.6)$ and $T_2 \sim \mathrm{Gomp}(a_2, b_2)$ with $(a_2, b_2) = (\log(0.01), 1)$, resulting in expected lifetimes equal to 3.6 and 4.1 respectively. Right-censoring is introduced via exponential distributed random variables with mean 10, resulting in a censoring of approximately 25%. We consider data of 5 different sample sizes,

$n = 30, 50, 100, 200, 500$ and, for each data set, we perform interval estimation of P_1 and P_2 at the time point $t^* = 4$.

Bayesian inference on $P_1(t^*)$ and $P_2(t^*)$ is based on 5000 posterior variates from $\pi(\theta \,|\, \text{data})$ and 5000 trajectories from the posterior beta process $A \,|\, \text{data}$. As for θ, we use the Jeffreys's prior in (3.2) and implement the adaptive rejection sampling coordinatewise for a total of 6000 iterations, discarding the first 1000 sweeps as burn-in. As for A, the prior process is centered at $A_0 = \alpha_0 t$ with $\alpha_0 = 0.165$ and we take $c(t) = m \exp(-\alpha_0 t)$ with $m = 1, 10, 20$, corresponding to three different degrees of prior beliefs. For each choice of m we compute credible intervals for $P_1(t^*)$ and $P_2(t^*)$ based on highest posterior density. Finally, we derive confidence intervals under the semiparametric model (1.1) based on the limiting normality of Corollary 2.2 and on plug-in estimates of the asymptotic variance (2.8). For comparison purposes, we also report interval estimation for the fully nonparametric case, where the cumulative CSH's are estimated via the Nelson-Aalen estimators, see ABGK Section IV.4.1.

In Table 1 we report the empirical frequentist coverage probability of the 95% interval estimates of $P_1(t^*)$ and $P_2(t^*)$ based on 1000 independent samples for each method (Bayesian with $m = 1, 10, 20$, semiparametric and nonparametric) and for each sample size ($n = 30, 50, 100, 200, 500$). Note that the performance of the Bayesian intervals increases for increasing sample size and decreasing concentration parameter m. A small coverage is associated with a big m because the initial guess on the cumulative overall hazard, i.e. A_0, is different from the true one: the expected lifetime corresponding to A_0 is approximately twice than $\mathrm{E}(T_1 \wedge T_2)$. As n increases, the data rule out the "wrong" prior guess more and more. The Bayesian intervals work well for all sample sizes for the least informative prior specification, which corresponds to $m = 1$: they attain coverage probabilities close to the nominal level and consistent with the ones of semiparametric estimation. The accuracy of intervals based on nonparametric estimation is somehow superior, even if one has to consider that the Bayesian intervals suffer of a wrong specification of the prior. Moreover, a better accuracy is also caused by the fact that the width of the nonparametric intervals is generally larger because of a larger asymptotic standard deviation.

6 Proofs of Theorem 2.1 and Theorem 4.1

Proof of Theorem 2.1. Result (2.5) is standard in survival analysis and holds under condition (A) and (B), see ABGK Section IV.1. Asymptotic normality of $\widehat{\theta}_n$ is derived using convex analysis in the spirit of Hjort and Pollard (1994). To this aim, the following Lemma is essential.

LEMMA 6.1 *Denote by E_j the index set for failures of the j-th type, $E_j = \{i : \ d_i = j, i = 1, \ldots, n\}$ and indicate by θ_j the two-dimensional vector $(\theta_{j1}, \theta_{j2})$, $j = 1, \ldots, k$. If for any $\theta =\in \mathbb{R}^{2k-2}$ there exists a triplet (i, j, h) with $j, h \in 1, \ldots, k, j \neq h, i \in E_j$ such that*

$$\langle (\theta_h - \theta_j), (1, t_i) \rangle > 0$$

Table 1: *Empirical coverage of interval estimates for $P_1(t^*)$ (upper block) and $P_2(t^*)$ (lower block) based on* 1000 *independent samples. Nominal coverage is* 95%.

	Method		**n = 30**	**n = 50**	**n = 100**	**n = 200**	**n = 500**
$\mathbf{P_1(t^*)}$	Bayesian	$(m = 1)$	.928	.936	.941	.940	.942
		$(m = 10)$	.905	.917	.933	.927	.937
		$(m = 20)$	.844	.864	.902	.911	.929
	Semiparametric		.916	.921	.927	.926	.929
	Nonparametric		.941	.946	.950	.946	.945
$\mathbf{P_2(t^*)}$	Bayesian	$(m = 1)$	.933	.933	.946	.939	.952
		$(m = 10)$	.899	.910	.934	.933	.942
		$(m = 20)$	.842	.863	.895	.915	.935
	Semiparametric		.901	.908	.921	.920	.929
	Nonparametric		.932	.944	.951	.944	.953

then the maximum likelihood estimate $\widehat{\theta}$ exists and is unique. Moreover, the likelihood has limit $-\infty$ at infinity and is strictly concave.

PROOF. The log likelihood $\ell_n(\theta)$ in (2.4) can be written as

$$\ell_n(\theta) = -\sum_{j\le k}\sum_{i\in E_j} \log\big\{\textstyle\sum_{l\le k} \exp\langle \theta_l - \theta_j, (1, t_i)^{\mathrm{t}}\rangle\big\}$$

Then, existence and uniqueness of the maximum likelihood estimate is implied by the overlap of observed failure times, see Albert and Anderson (1984). The condition of overlap has a simple geometric interpretation when $k = 2$, that is the two sets of observations $\{t_i,\ i \in E_1\}$ and $\{t_i,\ i \in E_2\}$ cannot be separated by any value on the time axis. □

We can now proceed with the proof of Theorem 2.1. Consider first the following Taylor expansion of $\log p_k(t, \theta)$ around θ_{tr}:

$$\log p_k(t, \theta_{\mathrm{tr}}) - \log p_k(t, \theta_{\mathrm{tr}} + x) = e(t, \theta_{\mathrm{tr}})^{\mathrm{t}} x + \frac{1}{2} x^{\mathrm{t}} V(t, \theta_{\mathrm{tr}}) x + \frac{1}{6} v(x, t, \theta_*)$$

for θ_* such that $|\theta_* - \theta_{\mathrm{tr}}| \le |x|$ and $v(x, t, \theta) = \sum_{j\le k} p_j(t, \theta)\{[z_j(t) - e(t, \theta)]^{\mathrm{t}} x\}^3$. A bound for the $v(x, t, \theta)$ can be found with techniques similar to those used by Hjort and Pollard (1994, Lemma A2). In fact, it can be shown that $|v(x, t, \theta)| \le 64(t^2 + 1)^{3/2}|x|^3$ regardless of the value of θ.

Next, Lemma 1 implies that, under condition (C), for a sufficiently large sample size the sequence of functions $C_n(x) = \ell_n(\theta_{\rm tr} + x/\sqrt{n}) - \ell_n(\theta_{\rm tr})$ is strictly concave in x with maximum in zero and can be written as

$$C_n(x) = H_n(\tau)^{\rm t} x - \frac{1}{2} x^{\rm t} \Sigma_n(\theta_{\rm tr}) x - \frac{1}{6} r_n(x, \theta_*), \tag{6.1}$$

where

$$H_n(t) = n^{-1/2} \textstyle\sum_{j \le k} \int_0^t [z_j(s) - e(s, \theta_{\rm tr})] {\rm d}N_j(s), \tag{6.2}$$

$$\Sigma_n(\theta) = n^{-1} \textstyle\int_0^\tau V(t, \theta) {\rm d}N_\cdot(t), \tag{6.3}$$

$$r_n(x, \theta_*) = n^{-3/2} \textstyle\int_0^\tau v(x, t, \theta_*) {\rm d}N_\cdot(t), \tag{6.4}$$

and θ_* in (6.4) such that $|\theta_* - \theta_{\rm tr}| \le |x|$. By the methods set forth in Section 1 of Hjort and Pollard (1994) convergence in (2.6) is implied by: (i) $r_n(x, \theta_*) \to_p 0$ for any θ_* such that $|\theta_* - \theta_{\rm tr}| \le |x|$, (ii) $H_n(\tau) \to_d {\rm N}(0, \Sigma)$ and (iii) $\Sigma_n(\theta_{\rm tr}) \to_p \Sigma$.

As for (i), $r_n(x, \theta_*)$ is bounded by $\int_0^\tau 64(t^2+1)^{3/2}|x|^3 n^{-3/2} {\rm d}N_\cdot(t)$, which is $O_p(n^{-1/2})$ by Condition (B) and an application of Lenglart's inequality. As for the proof of (ii) and (iii), we make use of convergence theory for counting processes. The martingale decomposition in (2.1) leads to

$$H_n(\tau) = n^{-1/2} \sum_{j \le k} \int_0^\tau [z_j(t) - e(t, \theta_{\rm tr})] {\rm d}M_j(t), \tag{6.5}$$

$$\Sigma_n(\theta_{\rm tr}) = n^{-1} \int_0^\tau V(t, \theta_{\rm tr}) Y(t) \alpha_{\rm tr}(t) {\rm d}t + n^{-1} \sum_{j \le k} \int_0^\tau V(t, \theta_{\rm tr}) {\rm d}M_j(t), \tag{6.6}$$

where $M_\cdot(t) = \sum_{j \le k} M_j(t)$. As for the convergence in distribution in (ii), note that $\{H_n(t),\ t \in [0, \tau]\}$ is a martingale so we can apply the Rebolledo's central limit theorem. The Lindeberg-type condition

$$n^{-1} \sum_{j \le k} \int_0^\tau |z_j(t) - e(t, \theta_{\rm tr})|^2 \, I\{n^{-1/2} |z_j(t) - e(t, \theta_{\rm tr})| \ge \epsilon\} Y(t) p_j(t, \theta_{\rm tr}) \alpha_{\rm tr}(t) {\rm d}t \to_p 0$$

is satisfied because the indicator goes to zero for large n. It is easy to see that

$$\langle H_n \rangle(\tau) = n^{-1} \sum_{j \le k} \int_0^\tau [z_j(t) - e(t, \theta_{\rm tr})]^{\otimes 2} Y(t) p_j(t, \theta_{\rm tr}) \alpha_{\rm tr}(t) {\rm d}t \to_p \Sigma,$$

which completes the proof of (ii). As for the convergence in probability in (iii), the first term in (6.6) converges to Σ by Condition (A) and boundedness of the integrand. The latter follows from Condition (B) and a bound on $V(t, \theta_{\rm tr})$ similar to the one used for $v(x, t, \theta)$. The second term in (6.6) is $O_p(n^{-1/2})$, which can be proved by combining Lenglart's inequality, the formula of stochastic integration with respect to martingales and Condition (B). This proves (iii). Finally, the asymptotic independence of $\widehat{A}_n$ and $\widehat{\theta}_n$ is easily verified by writing $\sqrt{n}(\widehat{\theta}_n - \theta_{\rm tr}) = \Sigma^{-1} H_n(\tau) + o_p(1)$ and

$$\sqrt{n}\,[\widehat{A}_n(t) - A_{\rm tr}(t)] = \sqrt{n} \int_0^t I\{Y(s) > 0\} Y(s)^{-1} {\rm d}M_\cdot(s) + o_p(1).$$

It can be shown that $H_n(t)$ is componentwise orthogonal with respect to the first term in the right hand side. The proof is then complete.

Proof of Theorem 4.1. It is easy to see that convergence in (4.1) is implied by Theorem 2 in Kim and Lee (2004), which holds under the hypotheses made on the prior parameters of the beta process. In order to prove (4.2) the following lemma is needed.

LEMMA 6.2 *Assume conditions (A)-(C) hold and that $\pi(\cdot)$ is positive and continuous at θ_{tr}. Then*

$$\int_{\mathbb{R}^{2k-2}} |g_n(x) - \phi(x)\pi(\theta_{\mathrm{tr}})|\,\mathrm{d}x \to_p 0,$$

where $g_n(x) = \exp\{\ell_n(\widehat{\theta}_n + x/\sqrt{n}) - \ell_n(\widehat{\theta}_n)\}\pi(\widehat{\theta}_n + x/\sqrt{n})$ and $\phi(x) = \exp\{-x^{\mathrm{t}}\Sigma x/2\}$.

PROOF. The proof goes along with the decomposition

$$\begin{aligned}\int_{\mathbb{R}^{2k-2}} \big|g_n(x) - \phi(x)\pi(\theta_{\mathrm{tr}})\big|\mathrm{d}x \le \int_{|x|\le K} \big|g_n(x) - \phi(x)\pi(\theta_{\mathrm{tr}})\big|\mathrm{d}x + \int_{K<|x|<\delta\sqrt{n}} g_n(x)\mathrm{d}x \\ + \int_{|x|\ge\delta\sqrt{n}} g_n(x)\mathrm{d}x + \int_{|x|>K} \phi(x)\pi(\theta_{\mathrm{tr}})\mathrm{d}x\end{aligned}$$

Denote the four integrals in the right-hand side by I_1, I_2, I_3 and I_4 respectively. Since I_4 can be set as small as needed (ϕ is concave with maximum at 0) it is sufficient to find, for given $\epsilon > 0$, two positive constants K and δ such that $\Pr\{I_j > \epsilon\} \to 0$ for $j = 1, 2, 3$.

For I_1 we use the third-order Taylor expansion

$$\ell_n(\widehat{\theta}_n + x/\sqrt{n}) - \ell_n(\widehat{\theta}_n) = -\frac{1}{2}x^{\mathrm{t}}\Sigma_n(\widehat{\theta}_n)x - \frac{1}{6}r_n(x, \theta_*), \tag{6.7}$$

for θ_* such that $|\theta_* - \widehat{\theta}_n| \le |x|/\sqrt{n}$. The remainder $r_n(x, \theta_*)$ is defined as in (6.4) and $\Sigma_n(\widehat{\theta}_n)$ is defined accordingly to (6.2). Next, for $\phi_n(x) = \exp\{-x^{\mathrm{t}}\Sigma_n(\widehat{\theta}_n)x/2\}$,

$$I_1 \le \pi(\theta_{\mathrm{tr}}) \int_{|x|\le K} |\phi_n(x) - \phi(x)|\,\mathrm{d}x + \int_{|x|\le K} |g_n(x) - \phi_n(x)\pi(\theta_{\mathrm{tr}})|\,\mathrm{d}x. \tag{6.8}$$

It can be shown, exploiting the continuity of $V(t, \cdot)$ and Lenglart's inequality, that $\Sigma(\widehat{\theta}_n) \to_p \Sigma$. Then, $\sup_{|x|\le K} |\phi_n(x) - \phi(x)| \to_p 0$ and the positivity of $\pi(\theta_{\mathrm{tr}})$ implies that the first term in (6.8) goes to zero in probability for any finite K. As for the second term in (6.8), we have

$$|g_n(x) - \phi_n(x)\pi(\theta_{\mathrm{tr}})| \le \phi_n(x)\pi(\theta_{\mathrm{tr}})\bigg[(1+\eta_1)\sup_{|x|\le K}\big|\exp\big\{-r_n(x,\theta_*)\big\} - 1\big| + \eta_1\bigg],$$

where $\eta_1 = \sup_{|x|\le K}\big|\pi(\widehat{\theta}_n + x/\sqrt{n})/\pi(\theta_{\mathrm{tr}}) - 1\big|$. Note that $\eta_1 \to_p 0$ because of $\pi(\widehat{\theta}_n + x/\sqrt{n}) \to_p \pi(\theta_{\mathrm{tr}})$ uniformly on $|x| \le K$ and the positivity of $\pi(\theta_{\mathrm{tr}})$. Moreover, reasoning as in the proof of Theorem 2.1, one finds that, for any $K > 0$ and

$|x| \leq K$, $\sup_{|x|\leq K} |\exp\{-r_n(x,\theta_*)\} - 1| \to_p 0$. Since $\int_{|x|\leq K} \phi_n(x)\mathrm{d}x$ is bounded in probability, we conclude that, for any $K > 0$, $\Pr\{I_1 > \epsilon\} \to 0$.

Regarding I_2, we use the following bound for the third order term of the Taylor expansion in (6.7):

$$\big|r_n(x,\theta_*)\big| \leq \delta x^{\mathrm{t}}x \int_0^\tau \frac{32}{3}(t^2+1)^{3/2} n^{-1} \mathrm{d}N_\cdot(t),$$

where the integral in the right hand side is a quantity bounded in probability (see condition (A) and (B) and use Lenglart's inequality). Hence, there exists $M > 0$ large enough such that $\Pr\big\{|r_n(x,\theta_*)| > \delta M\ x^{\mathrm{t}}x\big\} \to 0$. Next define (i) $B_n = \Sigma_n(\widehat{\theta}_n) - \Sigma$ and $b_n = (2k-2)^2 \max_{i,j=1,\ldots,2k-2} |B_{nij}|$ such that $x^{\mathrm{t}}B_n x \leq b_n\ h^{\mathrm{t}}h$; (ii) λ as the smallest eigenvalue of Σ, such that $x^{\mathrm{t}}\Sigma x \geq \lambda\ x^{\mathrm{t}}x$. Hence, we have $-\frac{1}{2}x^{\mathrm{t}}\Sigma_n(\widehat{\theta}_n)x \leq -\frac{1}{2}\big(\lambda - b_n\big)x^{\mathrm{t}}x$. Since $\lambda > 0$ (Σ is positive definite) and $b_n \to_p 0$, for some ϵ_1 sufficiently small there exists $\delta = \delta(\lambda, M, \epsilon_1)$ such that $\lambda + \delta M/3 - \epsilon_1 > 0$ and $\Pr\Big\{\lambda - b_n - 2\delta M < \epsilon_1\Big\} \to 0$. Fix now $\epsilon > 0$ and use expansion in (6.7) to get

$$\begin{aligned}\Pr\{I_2 > \epsilon\} \leq \Pr\left\{\sup_{|x|\leq\delta\sqrt{n}} \Big|\pi(\widehat{\theta}_n + x/\sqrt{n})/\pi(\theta_{\mathrm{tr}})\Big| > \eta_2\right\} + \Pr\Big\{|r_n(x,\theta_*)| > \delta M\ x^{\mathrm{t}}x\Big\} \\ + \Pr\big\{b_n > \lambda + \delta M/3 - \epsilon_1\big\} + \Pr\left\{\int_{K<|x|<\delta\sqrt{n}} e^{-\epsilon_1 x^{\mathrm{t}}x/2}\,\mathrm{d}x > \frac{\epsilon}{\eta_2 \pi(\theta_{\mathrm{tr}})}\right\},\end{aligned}$$

where $\eta_2 = \sup_{|x|\leq 2\delta} |\pi(\theta_{\mathrm{tr}} + x)/\pi(\theta_{\mathrm{tr}})|$. The first term in the right hand side goes to zero because $|\widehat{\theta}_n - \theta_{\mathrm{tr}}| \leq 2\delta$ eventually. Since we have already shown that the second and the third terms go to zero in probability, it is sufficient to choose K large enough such that $\int_{|x|>K} e^{-\epsilon_1 x^{\mathrm{t}}x/2}\,\mathrm{d}x \leq \epsilon/\eta_2\pi(\theta_{\mathrm{tr}})$. Then, there exist $K, \delta > 0$ such that $\Pr\{I_2 > \epsilon\} \to 0$.

As for I_3, first consider that

$$\int_{|x|\geq\delta\sqrt{n}} g_n(x)\,\mathrm{d}x \leq n^{k-1} \sup_{|\theta - \widehat{\theta}_n|\geq\delta} \exp\{\ell_n(\theta) - \ell_n(\widehat{\theta}_n)\}.$$

The set $\{\theta :\ |\theta - \widehat{\theta}_n| \geq \delta\}$ is eventually contained in $\{\theta :\ |\theta - \theta_{\mathrm{tr}}| \geq \delta/2\}$. Therefore, by the concavity of $\ell_n(\theta)$, it suffices to prove that

$$n^{k-1} \sup_{|\theta - \theta_{\mathrm{tr}}| = \delta/2} \exp\{\ell_n(\theta) - \ell_n(\widehat{\theta}_n)\} \to_p 0, \tag{6.9}$$

Reasoning as in the proof of Theorem 2.1, it is possible to show that $n^{-1}\big(\ell_n(\theta) - \ell_n(\theta_{\mathrm{tr}})\big) \to_p d(\theta)$ uniformly on compact set, where $d(\theta)$ is a strictly concave function with maximum at θ_{tr} equal to zero. Finally, consistency of $\widehat{\theta}_n$ leads to

$$n^{k-1} \sup_{|\theta - \theta_{\mathrm{tr}}| = \delta/2} \exp\{\ell_n(\theta) - \ell_n(\widehat{\theta}_n)\} \leq n^{k-1} \sup_{|\theta - \theta_{\mathrm{tr}}| = \delta/2} \exp\big\{n[o_p(1) + d(\theta)]\big\}.$$

Hence, (6.9) holds because $n^{k-1}e^{nd(\theta)} \to 0$. We conclude that, for any $\epsilon, \delta > 0$, $\Pr\{I_3 > \epsilon\} \to 0$, and the proof is complete. □

Now we are in the position to prove the second statement of Theorem 4.1 in a straightforward way. Denote by $f_n(\cdot)$ the posterior density of $\sqrt{n}(\theta - \widehat{\theta}_n)$. We aim at proving that $f_n(\cdot)$ converges in L_1-norm to a multivariate normal density with zero mean and covariance matrix Σ^{-1}. For $x = \sqrt{n}(\theta - \widehat{\theta}_n)$,

$$f_n(x) \propto \exp\{\ell_n(\widehat{\theta}_n + x/\sqrt{n})\}\pi(\widehat{\theta}_n + x/\sqrt{n})$$

so that $f_n(x) = g_n(x)/G_n$, for g_n defined in Lemma 6.2 and $G_n = \int g_n(x)\mathrm{d}x$. For $\phi(x) = \exp\{-x^{\mathrm{t}}\Sigma x/2\}$, it is then sufficient to show that g_n converges in L_1 to $\phi(h)\pi(\theta_{\mathrm{tr}})$. Hence, Lemma 2 completes the proof.

7 Treatment for general p_j

As we already pointed out, equation (1.2) corresponds to cause-specific conditional probabilities in the case of Gompertz-type CSH's, see equation (1.3). Other forms for p_j are equally plausible, e.g. one could start from a different family of hazard rates. This suggests a natural way to generalize (1.2), a convenient form being

$$\left\{ \begin{array}{rcll} p_j(t,\theta) & = & e^{\lambda(t,\theta_j)}p_k(t,\theta), & j = 1,\ldots,k-1 \\ p_k(t,\theta) & = & 1\Big/\Big\{\sum_{h\le k} e^{\lambda(t,\theta_h)}\Big\} & \end{array} \right. \tag{7.1}$$

For simplicity we keep θ_j as the two dimensional vector $(\theta_{j1}, \theta_{j2})$, even if this limitation is not strictly necessary. In order to ensure identifiability, we set θ_k to satisfy the constraint $\lambda(t,\theta_k) = 1$. This entails θ to be a $(2k-2)$-dimensional parameter. The function $\lambda(t,\theta_j)$ is assumed to be twice differentiable in θ_j. For example, starting with Weibull-type CSH, one can set $\lambda(t,\theta_j) = \log[\theta_{j1}t^{\theta_{j2}}]$, $\theta_{k1} = 1$ and $\theta_{k2} = 0$.

The arguments set forth in the previous sections can be adapted to (7.1) without serious efforts. First note that the part regarding the overall hazard remains exactly the same and that the conditions for the prior distributions on A and θ remain also the same. We next replace Conditions (B) and (C) of Section 2 in order to make Theorem 2.1 and Theorem 4.1 hold, see Section 6. To this aim, we need to introduce some additional notation for the first and second derivatives of $\lambda(t,\theta_j)$ with respect to θ:

$$z_j(t,\theta) = \frac{\partial}{\partial\theta}\lambda(t,\theta_j) \qquad \text{and} \qquad Z_j(t,\theta) = \frac{\partial^2}{\partial\theta\partial\theta^{\mathrm{t}}}\lambda(t,\theta_j).$$

We also redefine $e(t,\theta) = \sum_{j\le k} z_j(t,\theta)p_j(t,\theta)$ and $V(t,\theta)$ as

$$V(t,\theta) := \sum_{j\le k} p_j(t,\theta)\frac{\partial}{\partial\theta^{\mathrm{t}}}\big[z_j(t,\theta) - e(t,\theta)\big] = \sum_{j\le k} p_j(t,\theta)[z_j(t,\theta) - e(t,\theta)]^{\otimes 2}$$

The equality is due to the fact that the terms involving $Z_j(t,\theta)$ cancel out. Finally, the information matrix Σ remains defined as $\Sigma = \int_0^\tau V(t,\theta_{\mathrm{tr}})y(t)\alpha_{\mathrm{tr}}(t)\mathrm{d}t$. Condition (B) is then replaced by the assumption that, for any $\delta > 0$ and for any θ such that $|\theta - \theta_0| < \delta$,

(B1') $\displaystyle\int_0^\tau \left(\max_{j\le k} |z_j(t,\theta)| \right)^4 \alpha_{\rm tr}(t)\mathrm{d}t < \infty;$

(B2') $\displaystyle\int_0^\tau \left(\max_{j\le k} |Z_j(t,\theta)_{h,i}| \right)^4 \alpha_{\rm tr}(t)\mathrm{d}t < \infty$ for any $h, i = 1, \ldots, 2k-2$;

(B3') $\{\ell^{(3)}_{n,hil}(\theta)\} = O_p(n), \qquad h, i, l, = 1, \ldots, 2k-2,$

where, in (B3'), $\ell^{(3)}_n(\theta)$ denotes the array of the third derivatives of the log-likelihood of θ, now given by $\ell_n(\theta) = \sum_{j\le k} \int_0^\tau \big[\lambda(t,\theta_j) + \log p_k(t,\theta)\big]\mathrm{d}N_j(t)$. Condition (C) is replaced by

(C') $\ell_n(\theta)$ is strictly concave in all its domain.

The statement of Theorem 2.1 remains valid upon substitution of $z_j(t)$ with $z_j(t,\theta)$. As for the proof of the asymptotic normality of $\widehat{\theta}_n$, Condition (B1') is sufficient for the convergence of $H_n(\tau) \to_d \mathrm{N}(0,\Sigma)$ whereas condition (B2') guarantees that $\Sigma_n(\theta_{\rm tr}) \to_p \Sigma$. Condition (B3') is needed for the the remainder

$$r_n(x,\theta) = \frac{1}{\sqrt{n}} \sum_{h,i,l=1}^{2k-2} \frac{x_h x_i x_l}{6}\, \ell^{(3)}_{n,hil}(\theta)/n$$

being asymptotically negligible for θ in a neighborhood of $\theta_{\rm tr}$. The asymptotic independence of $\widehat{A}_n$ and $\widehat{\theta}_n$ follows from arguments similar to those used in the proof of Theorem 2.1. We stress that, in general, one needs to check Condition (C') case by case.

As for Theorem 4.1, one can prove asymptotic normality of the posterior distribution of θ by using Conditions (A), (B1'), (B2'), (B3') and (C') with techniques similar to those set forth in the proof of Theorem 4.1. The steps which deserve attention are the following: (i) the asymptotic negligibility of the remainder term $r_n(x,\theta)$ in a neighborhood of $\widehat{\theta}_n$; (ii) the bound in probability of $r_n(x,\theta)$ in a $\sqrt{n}$-neighborhood of $\widehat{\theta}_n$; (iii) the consistency of the observed information matrix $\Sigma(\widehat{\theta}_n)$ and (iv) the pointwise convergence of $n^{-1}\big(\ell_n(\theta) - \ell_n(\theta_{\rm tr})\big)$ to a strictly concave function with maximum at $\theta_{\rm tr}$ equal to zero. Condition (B3') is involved in points (i) and (ii), whereas point (iii) is dealt with Condition (B1') and (B2') in conjunction with Lenglart's inequality. Finally, Condition (C') is the key ingredient for dealing with point (iv).

Remark 1 It is worth considering a subclass of models in (7.1) that arises when $\lambda(t,\theta_j)$ has the following form:

$$\lambda(t,\theta_j) = \exp\{\theta_{j1} + \theta_{j2} g(t)\},$$

with g is a continuous function defined on $\mathbb{R}_+$. It turns out that the arguments used for model (1.2) extend exactly in the same way if we replace the failure times $T_1, \ldots, T_n$ with $g(T_1), \ldots, g(T_n)$. In fact, $\lambda(t,\theta_j)$ has null derivatives of order higher than one and $z_j(t)$ has $(1, g(t))^{\rm t}$ in the j-th block and zeros elsewhere. Condition (B) is replaced

by assuming that $\int_0^\tau [g(t)^2+1]^3 \alpha_{\rm tr}(t)\mathrm{d}t < \infty$, whereas the strict concavity of $\ell_n(\theta)$ is assured by the overlap of the transformed times $g(T_1), \ldots, g(T_n)$. In particular, for monotonic g, overlap of the actual observed failure times is enough, so that Condition (C) remains the same. Note that $g(t) = \log(t)$ corresponds to the Weibull case, implying that our results cover this noteworthy case as well. □

8 Concluding remarks

In this paper we proposed novel Bayesian methods for the analysis of competing risks data and investigated the asymptotic normality of the posterior distribution of differentiable functionals of model parameters. The semiparametric formulation takes the cumulative overall hazard as the infinite-dimensional parameter and is characterized by the factorization in the likelihood. This simplifies the derivation of the BvM theorem for differentiable functionals, since the joint posterior distribution of (A, θ) is the product of the marginals under independent priors and it is sufficient to prove the BvM for A and θ separately. This suggests that the way the model is formulated matters in order to exploit the functional delta method. On the other side, model (1.1) is appealing since the cause-specific conditional probability p_j is a primary result of estimation. Indeed, these conditional probabilities are of direct statistical interest in many practical situations: a simultaneous plot of the p_j's against time might serve as an additional graphical device for describing the prevalence of risks, which is a central topic in many applications. A common approach for studying of prevalence of risks is via the CIF's P_j, even though they do not provide changes in the relative risk of failure, which is rather given by the CSH's. The quantity p_j is a valid alternative, presenting the advantage of having a sound interpretation as conditional probability.

The competing risks problem is a special case when survival times are associated to mark variables which are not observed when the event is censored. Huang and Louis (1998) discuss nonparametric methods in this framework: in particular they propose to estimate the joint distribution of survival times and mark variables by using the concept of cumulative mark-specific hazard function. A natural candidate for Bayesian estimation of this type of models is represented by the family of spatial neutral to the right processes, recently introduced by James (2006). It will be then of interest to study the asymptotic properties of the posterior distribution in this setting.

References

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *Ann. Statist.* **6**, 701–726.

Albert, A. and Anderson, J.A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **71**, 1–10.

Andersen, P.K., Borgan, Ø., Gill, R.D. and Keiding, N. (1993). *Statistical models based on counting processes.* Springer, New York.

Cox, D.D. (1993). An analysis of Bayesian inference for nonparametric regression. *Ann. Statist.* **21**, 903–923.

De Blasi, P. and Hjort, N.L. (2007). Bayesian survival analysis in proportional hazard models with logistic relative risk. *Scand. J. Statist.* **34**, 229–257.

Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates. *Ann. Statist.* **14**, 1–26.

Doksum, K. (1974). Tailfree and neutral random probabilities and their posterior distributions. *Ann. Probab.* **2**, 183–201.

Freedman, D. (1999). On the Bernstein-von Mises theorem with infinite-dimensional parameters. *Ann. Statist.* **27**, 1119–1140.

Gasbarra, D. and Karia, S.R. (2000). Analysis of competing risks by using Bayesian smoothing. *Scand. J. Statist.* **27**, 605-617.

Gilks, W.R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Appl. Statist.* **41**, 337–348.

Gill, R.D. (1989). Non- and semi-parametric maximum likelihood estimators and the Von Mises method I. *Scand. J. Statist.* **16**, 97–128 (with discussion).

Gill, R.D. and Johansen, S. (1990). A survey of product-integration with a view toward application to survival analysis. *Ann. Statist.* **18**, 1501–1555.

Hjort, N.L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *Ann. Statist.* **18**, 1259–1294.

Hjort, N.L. and Pollard, D.B. (1994). Asymptotics for minimisers of convex processes. *Statistical Research Report*, Department of Mathematics, University of Oslo.

Huang, Y. and Louis, T.A. (1998). Nonparametric estimation of the joint distribution of survival time and mark variables. *Biometrika* **85**, 785–798.

Ibrahim, J.G. and Laud, P.W. (1991). On Bayesian analysis of generalized linear models using Jeffreys's prior. *J. Amer. Statist. Assoc.* **86**, 981–986.

James, L.F. (2006). Poisson calculus for spatial neutral to the right processes. *Ann. Statist.* **34**, 416–440.

Kim, Y. (1999). Nonparametric Bayesian estimators for counting processes. *Ann. Statist.* **27**, 562–588.

Kim, Y. (2006). The Bernstein-von Mises theorem for the proportional hazard model. *Ann. Statist.* **34**, 1678–1700.

Kim, Y. and Lee, J. (2004). A Bernstein–von Mises theorem in the nonparametric right-censoring model. *Ann. Statist.* **32**, 1492–1512.

Larson, M.G. and Dinse, G.E. (1985). A mixture model for the regression analysis of competing risks data. *Appl. Statist.* **34**, 201–211.

Lawless, J.F. (2003). *Statistical models and methods for lifetime data. Second edition.* John Wiley and Sons.

Salinas-Torres, V.H., Pereira, C.A.B. and Tiwari, R.C. (2002). Bayesian nonparametric estimation in a series system or a competing-risks model. *Nonparametric Statistics* **14**, 449–458.

Sato K. (1999). *Lévy Processes and Infitely Divisible Distributions*. Cambridge University Press.

Schervish, M.J. (1995). *Theory of statistics*. Springer, New York.

Zhao, L.H. (2000). Bayesian aspects of some nonparametric problems. *Ann. Statist.* **28**, 532–552.