Journal of Statistical Planning and Inference 139 (2009) 2316 – 2328

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

The Bernstein–von Mises theorem in semiparametric competing risks models

Pierpaolo De Blasi[a,*,1], Nils Lid Hjort[b]

[a] *Dipartimento di Statistica e Matematica Applicata, Università degli Studi di Torino, via Maria Vittoria 38, 10100 Torino, Italy*
[b] *Matematisk Institutt, Universitet i Oslo, PO Box 1053, Blindern 0316 Oslo, Norway*

ARTICLE INFO

Article history:
Received 30 April 2008
Received in revised form
23 October 2008
Accepted 24 October 2008
Available online 31 October 2008

MSC:
62F15
62G20

Keywords:
Bayesian nonparametrics
Bernstein–von Mises theorem
Beta process
Cause-specific conditional probability
Competing risks
Semiparametric inference

ABSTRACT

Semiparametric Bayesian models are nowadays a popular tool in event history analysis. An important area of research concerns the investigation of frequentist properties of posterior inference. In this paper, we propose novel semiparametric Bayesian models for the analysis of competing risks data and investigate the Bernstein–von Mises theorem for differentiable functionals of model parameters. The model is specified by expressing the cause-specific hazard as the product of the conditional probability of a failure type and the overall hazard rate. We take 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 first develop the large-sample properties of maximum likelihood estimators by giving simple sufficient conditions for them to hold. Then, we show that, under the chosen priors, 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.

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 consistency and asymptotic normality of the MLE. See, e.g., Ghosh and Ramamoorthi (2003) 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. See Cox (1993), Freedman (1999) and 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

* Corresponding author.
E-mail addresses: pierpaolo.deblasi@unito.it (P. De Blasi), nils@math.uio.no (N.L. Hjort).
[1] Also affiliated to Collegio Carlo Alberto.

0378-3758/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2008.10.018

stationary increments. See, e.g., Sato (1999). This approach was first introduced by Hjort (1990) and developed further by Kim (1999). Kim and Lee (2004) proceed by proving that the BvM theorem holds for the cumulative hazard function in terms of the large-sample distribution of the Nelson–Aalen estimator and, then, extend the result to the survival function via the functional delta method. The survival function is in fact recovered from the cumulative hazard via the product integral, which is compactly differentiable. 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 carried out by Kim (2006) and De Blasi and Hjort (2007). Their results witness the convenience of working at the cumulative hazard level, in that Bayesian methods can be readily extended to more complex event history data by using the multiplicative intensity model of Aalen (1978).

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. We adopt a semiparametric formulation and consider frequentist estimation by following a counting process approach, see the monograph by Andersen et al. (1993) (ABGK henceforth). The quantities of interest are expressed as differentiable functionals of the model parameters, say (A, θ), where A is the functional parameter and θ is the finite-dimensional parameter. A property of the considered formulation is that the total likelihood factorizes in two parts, corresponding to A and θ, so that the derivation of the large-sample properties of the corresponding MLE is simplified. The functional delta method is then the key tool for deriving the limiting distribution of the functionals of interest. We exploit the factorization in the likelihood in order to establish a BvM theorem for the aforementioned functionals: the strategy is to take A and θ independent in the prior in order to get independence in the posterior; thus, the asymptotic normality follows from a further application of the functional delta method, provided that we establish the BvM theorem for the marginal posterior distributions of A and θ. As for the infinite-dimensional parameter, we utilize the results in Kim and Lee (2004) by appropriately choosing the prior. The main effort for θ consists in establishing the good behavior of its conditional likelihood: the continuity and nonnegativity of the prior density leads then to the BvM theorem by using arguments similar to those in Ghosh and Ramamoorthi (2003, Theorem 1.4.2). We are then able to provide simple and neat sufficient conditions for the BvM theorem to hold for any differentiable functional of the model parameters.

2. Semiparametric competing risks models

The semiparametric model has statistical interest on its own and is described as follows. We consider the pair (T^0, D^0), where T^0 is the failure time and $D^0 \in \{1, \dots, k\}$ is the type of failure, meaning 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:

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

that is $\alpha_j(t)$ is the instantaneous rate of a failure of type j at time t. Alternatively, one can describe competing risk data via the marginal probabilities $P_j(t) = \Pr\{T^0 \leqslant t, D^0 = j\}$, which are known as cumulative incidence functions (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, while, in general, it is a right-continuous, nondecreasing function with jumps in $(0, 1)$.

We work in a semiparametric setting and specify A_j as follows:

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

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 $\sum_{j \leqslant k} p_j(s, \theta) = 1$ for each s and θ; (ii) $A = \sum_{j \leqslant k} A_j$ is the cumulative overall hazard and is left unspecified. Our results apply to a general class of models, for p_j as detailed in Section 8 (see Eq. (26) therein). However, in order to ease the flow of ideas and avoid heavy notation, we focus first on the crucial case

$$\begin{cases} p_j(s, \theta) = \mathrm{e}^{\theta_{j1} + \theta_{j2} s} p_k(s, \theta), & j = 1, \dots, k-1, \\ p_k(s, \theta) = \left\{ \sum_{h \leqslant k} \mathrm{e}^{\theta_{h1} + \theta_{h2} s} \right\}^{-1}, \end{cases} \tag{2}$$

where $\theta_{k1} = \theta_{k2} = 0$ 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:

$$p_j(t, \theta) = \lim_{\Delta t \downarrow 0} \frac{\Pr\{T^0 \in [t, t + \Delta t), D^0 = j\}}{\Pr\{T^0 \in [t, t + \Delta t)\}} = \frac{\mathrm{d}A_j(t)}{\mathrm{d}A(t)}, \tag{3}$$

see ABGK, Section II.6. Following the terminology of Gasbarra and Karia (2000), we call p_j the *cause-specific conditional probability*.

Note that p_j corresponds to the subdensity $\mathrm{d}P_j/\mathrm{d}t$ normalized over its sum, while, in the continuous case, it is recovered from the CSHs via normalization, i.e. $p_j(t)=\alpha_j(t)/\sum_{h\leqslant k}\alpha_h(t)$. In particular, (2) can be seen as the cause-specific conditional probabilities that arise by starting with Gompertz-type CSH and by normalization for the k-th one. The cause-specific conditional probability describes the relative risk of a failure type in terms of the overall rate of failure as it varies over time. The common approach to the study of prevalence of risks is via the CIFs, even though they do not provide changes in the relative risk of failure. The quantity p_j is a valid alternative, presenting the advantage of having a sound interpretation as conditional probability. There exist other quantities that can be used to describe competing risk data, depending on the nature of the application at issue, see the discussion in Pepe and Mori (1993). In order to avoid confusion, the reader should note the difference with the conditional probability $\Pr(T^0\leqslant t|D^0=j)$, sometimes used for describing competing risks data in the so-called mixture model. See, e.g., Larson and Dinse (1985).

Estimation of the joint distribution of (T^0,D^0) is carried out via the cause-specific conditional probability p_j and the cumulative overall hazard A. The model parameters (A,θ) take values 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 functions 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)), 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

$$P_j(t)=\int_0^t\prod_{[0,s)}\{1-\mathrm{d}A(u)\}p_j(s,\theta)\,\mathrm{d}A(s),\quad j=1,\dots,k. \tag{4}$$

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 facts: (i) for A_j given in (1), the mapping $\Psi:(A,\theta)\to(A_1,\dots,A_k)$ 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). Thus, it follows that S and P_j are compactly differentiable functionals of (A,θ) as well.

The rest of the paper is organized as follows. In Section 3, we introduce the counting process formulation of cause-specific failure times and show that a factorization in the likelihood of (A,θ) takes place. We exploit the similarity of $p_j(t,\theta)$ in (2) with a multinomial logistic regression model in order to derive the asymptotic properties of the MLE for θ, while nonparametric estimation for A is obtained via the Nelson–Aalen estimator. Then, the allied limiting distributions for A_j, S and P_j are derived using the functional delta method. In Section 4, we develop Bayesian methods by using a Jeffreys prior for θ and specifying a nonparametric prior for A via the beta process of Hjort (1990). Section 5 provides the two BvM theorems for A and θ, which, in turns, imply the asymptotic normality of the posterior distribution of A_j and P_j. In Section 6, a simulation study is presented, with focus on the empirical coverage of credible sets for $P_j(t)$. In order to ease the flow of ideas, we collect the more technical proofs in Section 7. In Section 8, we provide conditions for the BvM theorem to hold for a larger family of models which includes (2) as a special case; we also show how the previously developed techniques may be adapted to this family. Finally, some concluding remarks and lines of future research are provided.

3. 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 $(T_1,D_1),\dots,(T_n,D_n)$, where T_i is the (possibly right-censored) failure time and $D_i=\{0,1,\dots,k\}$ 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. For each observation (T_i,D_i), we consider k counting processes $N_{i,j}(t)=I\{T_i\leqslant t,D_i=j\}$, $j=1,\dots,k$, and the at-risk process $Y_i(t)=I\{T_i\geqslant t\}$. According to Aalen's multiplicative intensity model, $N_{i,j}$ can be decomposed as

$$N_{i,j}(t)=\int_0^t Y_i(s)\,\mathrm{d}A_j(s)+M_{i,j}(t),\quad j=1,\dots,k, \tag{5}$$

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 (5), as well as orthogonality, and will be denoted by suppressing the index i. We also make use of the aggregated process $N_{\cdot}(t)=\sum_{j\leqslant k}N_j(t)$, 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 multinomial form:

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

cf. Eq. (2.7.2″) in ABGK, Section II.7. Upon substitution of (1), the part involving θ can be separated out from that involving A, so that $L(A,\theta)$ factorizes in two components, say $L(A)$ and $L(\theta)$, where

$$L(A)=\prod_t \left\{ \mathrm{d}A(t)^{\mathrm{d}N_\cdot(t)} \{1-\mathrm{d}A(t)\}^{Y(t)-\mathrm{d}N_\cdot(t)} \right\}, \tag{6}$$

$$L(\theta)=\prod_{i\leqslant n}\prod_{j\leqslant k} p_j(\theta,t_i)^{\Delta N_{i,j}(t_i)}. \tag{7}$$

Here $L(A)$ is the likelihood of the cumulative overall hazard, and simply corresponds to the case of right-censored survival times; $L(\theta)$ is the conditional likelihood of θ and depends only on the uncensored observations. The factorization of $L(A,\theta)$ represents an important property of model (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 the MLE $\widehat{\theta}_n$ for θ. Note that the Nelson–Aalen estimator has a maximum likelihood interpretation, see ABGK, Section IV.1. Upon substitution of (2) for $p_j(\theta,t)$ in (7), the conditional likelihood $L(\theta)$ is similar to 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 the good properties of the multinomial logistic regression model, in particular the strict concavity of the log-likelihood of the regression coefficient provided that there is overlap on the space of covariate variables, see Albert and Anderson (1984). This allows us to provide simple and neat conditions for the large-sample properties of $\widehat{\theta}_n$ to hold.

To this aim, we postulate that model (1)–(2) is in force for a certain cumulative overall hazard A_{tr}, with derivative α_{tr}, 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 required conditions are the following:

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

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. The infinite interval case cannot be derived from the finite interval case, since one has $\int_0^\infty \alpha_{\mathrm{tr}}(t)\,\mathrm{d}t=\infty$. Some extra conditions have then to be made on the likelihood contribution in the tail (τ,∞), see, e.g., Andersen and Gill (1982, Theorem 4.2). Condition (C) assures that, for sufficiently large sample sizes, there is overlap of cause-specific failure times, which is needed for the concavity of $\log L(\theta)$, see Lemma 1 in Section 7.

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

$$\ell_n(\theta)=\sum_{j\leqslant k}\int_0^\tau [z_j(t)^{\mathrm{t}}\theta+\log\, p_k(t,\theta)]\,\mathrm{d}N_j(t), \tag{8}$$

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 $e(t,\theta)=\sum_{j\leqslant k} z_j(t)p_j(t,\theta)$, 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 $V(t,\theta)=\sum_{j\leqslant k}p_j(t,\theta)[z_j(t)-e(t,\theta)]^{\otimes 2}$, 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 $b=(b_1,\ldots,b_p)^{\mathrm{t}}$, write $|b|=(b^{\mathrm{t}}b)^{1/2}$ and $b^{\otimes 2}=bb^{\mathrm{t}}$, 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 j-th block equal to the j-th column of B. The covariance matrix of a matrix-valued random variable $\mathbf{X}$ is then defined as $\mathrm{cov}(\mathbf{X})=\mathrm{E}\{[\mathrm{vec}(\mathbf{X})-\mathrm{vec}(\mathrm{E}\mathbf{X})][\mathrm{vec}(\mathbf{X})-\mathrm{vec}(\mathrm{E}\mathbf{X})]^{\mathrm{t}}\}$. Finally, let $C_{[0,\tau]}$ be the space of continuous functions defined on the interval $[0,\tau]$.

Theorem 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*

$$\sqrt{n}(\widehat{A}_n-A_{\mathrm{tr}})\to_{\mathrm{d}}\mathrm{W}(\sigma^2)\quad \text{on } D_{[0,\tau]}, \tag{9}$$

$$\sqrt{n}(\widehat{\theta}_n-\theta_{\mathrm{tr}})\to_{\mathrm{d}}\mathrm{N}(0,\Sigma^{-1}). \tag{10}$$

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. 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 Eqs. (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 $\Psi_1(\theta) = [p_1(t,\theta), \ldots, p_k(t,\theta)]^{\mathrm{t}}$ 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 integral (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 provides 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) 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 1. *For $j = 1, \ldots, k$ define the Gaussian process*

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

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

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

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 $0 \leqslant s \leqslant t$ and write $\mathbf{P}(t) = \mathbf{P}(0,t)$. Then

$$\mathbf{P}(t) = \prod_{[0,t]} \{\mathbf{I} + \mathrm{d}\mathbf{A}(s)\}, \tag{11}$$

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 through $\{1, \ldots, k\}$. Note that $[S(t), P_1(t), \ldots, P_k(t)]$ corresponds to the first row of $\mathbf{P}(t)$. The thesis of Corollary 1 can be written as $\sqrt{n}(\widehat{\mathbf{A}} - \mathbf{A}) \to_{\mathrm{d}} \mathbf{U}$ 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 $\mathrm{cov}(\mathbf{U}(t)) = \sum_{j,h \leqslant k} \mathrm{vec}(C_j)\,\mathrm{vec}(C_h)^{\mathrm{t}} \omega_{jh}(t)$, where

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

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 for $\mathbf{A}$ in (11) via Eq. (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. *The transition matrix $\widehat{\mathbf{P}}$ has the following asymptotic behavior:*

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

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 given by

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

with asymptotic variance

$$\mathrm{var}(Z_j(t)) = \int_0^t S(u-)^2 \left\{ \frac{(P_j(t) - P_j(u))^2}{S(u)^2}\sigma^2(u)\,\mathrm{d}u - 2\frac{P_j(t) - P_j(u)}{S(u)} p_j(u, \theta_{\mathrm{tr}})\sigma^2(u)\,\mathrm{d}u + \mathrm{d}\omega_{jj}(u) \right\}. \tag{12}$$

4. Bayesian inference

As far as model (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 in the prior and, then, derive the posterior distributions separately because of the factorization in $L(A, \theta)$. As for the prior distribution of A, we employ the beta process of Hjort (1990), which is an increasing additive process with paths that lie, with probability one, in the space of cumulative hazard functions. A formal definition of the beta process is as follows. Let $\mathscr{A}$ is the set of right-continuous nondecreasing functions A on $\mathbb{R}_+$, such that $A(0) = 0$ and A is increasing to infinity with jumps in $(0, 1)$. Let $A_0 \in \mathscr{A}$ with jumps at points $\{t_1, t_2, \ldots\}$ and let c be a piecewise continuous, nonnegative real valued function on $[0, \infty)$. Then, a beta process with parameters c and A_0, in symbols $A \sim \mathrm{Beta}(c, A_0)$, is defined as an increasing additive process with Lévy–Khinchine representation given by

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

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 $A_{0,c}(t) = A_0(t) - \sum_{t_j \leqslant t} \Delta A_0(t_j)$. 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 $1 - \prod_{[0,t]}\{1 - \mathrm{d}A(s)\}$ distributed as a Dirichlet process with prior mean equal to $[1 - \prod_{[0,t]}\{1 - \mathrm{d}A_0(s)\}]$ 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 is 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 Eq. (6)). Then, one can exploit the conjugacy of the beta process to i.i.d. right-censored lifetimes to get

$$A|\mathrm{data} \sim \mathrm{Beta}\left\{ c + Y, \int \frac{c\mathrm{d}A_0 + \mathrm{d}N_\cdot}{c + Y} \right\}, \tag{13}$$

where we have used the counting process notation introduced in Section 2. The second parameter in (13) 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

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

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 (13).

As for the parameter θ, any density function $\pi(\theta)$ with support $\mathbb{R}^{2k-2}$ can be used, although specifying a meaningful distribution can be a difficult task. A possibility is to adopt the prior density dictated by Jeffreys rule, which is suited to the case of little available prior information. For general parametric models, the Jeffreys 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 prior is given by

$$\pi(\theta) \propto \left| \sum_{i:d_i \neq 0} V(t_i, \theta) \right|^{1/2}. \tag{14}$$

See 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 log-concavity of $L(\theta)$. In fact, by Lemma 1 in the Section 6.1, $L(\theta)$ is log-concave as long as there is overlapping of cause-specific failure times. In this case, a log-concave prior $\pi(\theta)$ leads to a log-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 CIFs can be obtained via posterior averaging. The trajectory $\{A^*(t), t \geqslant 0\}$ will be of pure jumps, with jump times $0 < s_1 < s_2 < \cdots$ obtained by reordering the uncensored failure times in the data and the time points used in the discretization of the continuous part. Then, we compute

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

as a variate from the posterior distribution of P_j, see Eq. (4).

It is worth mentioning that a related Bayesian semiparametric approach to the analysis of competing risks data can be found in Gasbarra and Karia (2000). They 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 to each segment an independent k-dimensional Dirichlet distributed random vector. Then p_j is obtained by kernel smoothing. As for the overall hazard, they use a convolution of gamma processes at the hazard rate level, in the spirit of Lo and Weng (1989). A fully nonparametric treatment of the competing risks model is instead implicit in Hjort (1990, Section 5), where the beta process is used for Bayesian inference on nonhomogeneous Markov processes. The idea is to assign independent beta processes to all CSHs, using the fact that, given a sample of right-censored cause-specific failure times, they preserve independence and are still beta distributed in the posterior.

5. BvM theorem

In this section we provide the BvM theorem for the posterior distribution of A_j and P_j. Indeed, we prove that, under a beta process prior 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 $(\widehat{A}_n, \widehat{\theta}_n)$, as stated in Theorem 1. The claimed BvM result for A_j and P_j is, then, a direct consequence of the functional delta method. Note that the same is true for all differentiable functionals of (A, θ) or of the CSHs 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)–(2) with true parameters $(A_{\text{tr}}, \theta_{\text{tr}})$, 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 2. *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 $0 < \inf_{t\in[0,\tau]} c(t) \leqslant \sup_{t\in[0,\tau]} c(t) < \infty$. Moreover, assume π to be positive and continuous at θ_{tr}. Then, as $n \to \infty$,*

$$\sqrt{n}(A - \widehat{A}_n)|\text{data} \to_{\text{d}} \text{W}(\sigma^2) \quad \textit{on } D_{[0,\tau]} \textit{ in probability}, \tag{15}$$

$$\sqrt{n}(\theta - \widehat{\theta}_n)|\text{data} \to_{\text{d}} \text{N}(0, \Sigma^{-1}) \quad \textit{in probability}. \tag{16}$$

Convergence in (15) is implied by Theorem 2 in Kim and Lee (2004), which holds under the hypotheses made on the prior parameters c and A_0 of the beta process. The proof of (16) is deferred to Section 6, where we extend the arguments in Theorem 1.4.2 in Ghosh and Ramamoorthi (2003) to the multiparameter case and adapt them to the present setting.

The independence in the posterior distribution of A and θ implies that a BvM theorem also holds for their joint distribution in accordance with the asymptotic result in Theorem 1. Corollaries 3 and 4 provide the BvM for A_j and P_j and follow from applications of the functional delta method, as implemented in Section 2. See also Kim and Lee (2004).

Corollary 3. *Under the hypotheses of Theorem* 2

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

on $(D_{[0,\tau]})^k$ in probability, where $\widehat{A}_j(t)$ and $U_j(t)$ are defined as in Corollary 1.

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

Method	$n=30$	$n=50$	$n=100$	$n=200$	$n=500$
$P_1(t^*)$					
Bayesian					
$(m=1)$	0.928	0.936	0.941	0.940	0.942
$(m=10)$	0.905	0.917	0.933	0.927	0.937
$(m=20)$	0.844	0.864	0.902	0.911	0.929
Semiparametric	0.916	0.921	0.927	0.926	0.929
Nonparametric	0.941	0.946	0.950	0.946	0.945
$P_2(t^*)$					
Bayesian					
$(m=1)$	0.933	0.933	0.946	0.939	0.952
$(m=10)$	0.899	0.910	0.934	0.933	0.942
$(m=20)$	0.842	0.863	0.895	0.915	0.935
Semiparametric	0.901	0.908	0.921	0.920	0.929
Nonparametric	0.932	0.944	0.951	0.944	0.953

Nominal coverage is 95%.

Corollary 4. *Under the hypotheses of Theorem* 2, *for* $j=1,\dots,k$,

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

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

6. Simulation study

In this section we investigate the validity of the BvM theorem on simulated data by comparing the coverage probability of frequentist 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) with Gompertz distributions, so that model (1)–(2) is in force. Specifically, $T_1\sim\text{Gomp}(a_1,b_1)$ with $(a_1,b_1)=(\log(0.05),0.6)$ and $T_2\sim\text{Gomp}(a_2,b_2)$ 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 times with mean 10, resulting in a censoring of approximately 25%. We consider data of five 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 prior in (14) and implement the adaptive rejection sampling coordinate-wise for a total of 6000 iterations, discarding the first 1000 sweeps as burn-in. As for A, the prior process is centered at $A_0(t)=\alpha_0 t$ 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) based on the limiting normality of Corollary 2 and the plug-in estimates for the asymptotic variance in (12). For comparison purposes, we also report interval estimation for the fully nonparametric case, where the cumulative CSHs 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 as big as $\mathrm{E}(T_1\wedge T_2)$. 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 coverage of intervals based on nonparametric estimation is generally superior, although one has to consider that the Bayesian intervals suffer from the specification of the prior beliefs on A. Moreover, the nonparametric intervals show a better coverage than the semiparametric ones, which might be due to the approximation involved in the derivation of the asymptotic variance via the functional delta method.

7. Proofs of results in Sections 2 and 4

7.1. Proof of Theorem 1

Result (9) is standard in survival analysis and holds under Conditions (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 1. *Denote by E_j the index set for failures of the j-th type, $E_j = \{i : d_i = j, i = 1, \dots, n\}$ and indicate by θ_j the j-th block of θ, i.e. $\theta_j = (\theta_{j1}, \theta_{j2})^{\mathrm{t}}, j = 1, \dots, k-1$ and $\theta_k = (0,0)^{\mathrm{t}}$. If for any $\theta \in \mathbb{R}^{2k-2}$ there exists a triplet (i,j,h) with $j,h \in 1, \dots, k,\ j \neq h, i \in E_j$ such that*

$$(\theta_{h1} - \theta_{j1}) + t_i(\theta_{h2} - \theta_{j2}) > 0$$

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

Proof. The log-likelihood $\ell_n(\theta)$ in (8) can be written as

$$\ell_n(\theta) = -\sum_{j \leqslant k} \sum_{i \in E_j} \log \left\{ \sum_{l \leqslant k} \exp[(\theta_{l1} - \theta_{j1}) + t_i(\theta_{l2} - \theta_{j2})] \right\}.$$

Then, existence and uniqueness of the MLE 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 now proceed with the proof of (10). Consider the following Taylor expansion of $\log p_k(t,\theta)$ around θ_{tr}:

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

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

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

$$C_n(x) = H_n(\tau)^{\mathrm{t}} x - \tfrac{1}{2} x^{\mathrm{t}} \Sigma_n(\theta_{\mathrm{tr}}) x - \tfrac{1}{6} r_n(x,\theta_*), \tag{17}$$

where

$$H_n(t) = n^{-1/2} \sum_{j \leqslant k} \int_0^t [z_j(s) - e(s,\theta_{\mathrm{tr}})]\, \mathrm{d}N_j(s), \tag{18}$$

$$\Sigma_n(\theta) = n^{-1} \int_0^\tau V(t,\theta)\, \mathrm{d}N_\cdot(t), \tag{19}$$

$$r_n(x,\theta_*) = n^{-3/2} \int_0^\tau \nu(x,t,\theta_*)\, \mathrm{d}N_\cdot(t), \tag{20}$$

and θ_* such that $|\theta_* - \theta_{\mathrm{tr}}| \leqslant |x|$. By the methods set forth in Section 1 of Hjort and Pollard (1994), convergence in (10) is implied by: (i) $r_n(x,\theta_*) \to_{\mathrm{p}} 0$ for any θ_* such that $|\theta_* - \theta_{\mathrm{tr}}| \leqslant |x|$, (ii) $H_n(\tau) \to_{\mathrm{d}} \mathrm{N}(0,\Sigma)$ and (iii) $\Sigma_n(\theta_{\mathrm{tr}}) \to_{\mathrm{p}} \Sigma$.

As for (i), $r_n(x,\theta_*)$ is bounded by $\int_0^\tau 64(t^2+1)^{3/2}|x|^3 n^{-3/2}\, \mathrm{d}N_\cdot(t)$, which is $\mathrm{O_p}(n^{-1/2})$ 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 (5) leads to

$$H_n(\tau) = n^{-1/2} \sum_{j \leqslant k} \int_0^\tau [z_j(t) - e(t,\theta_{\mathrm{tr}})]\, \mathrm{d}M_j(t), \tag{21}$$

$$\Sigma_n(\theta_{\mathrm{tr}}) = n^{-1} \int_0^\tau V(t,\theta_{\mathrm{tr}}) Y(t) \alpha_{\mathrm{tr}}(t)\, \mathrm{d}t + n^{-1} \sum_{j \leqslant k} \int_0^\tau V(t,\theta_{\mathrm{tr}})\, \mathrm{d}M_j(t), \tag{22}$$

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

$$n^{-1} \sum_{j \leqslant k} \int_0^\tau |z_j(t) - e(t,\theta_{\mathrm{tr}})|^2\, I\{n^{-1/2}|z_j(t) - e(t,\theta_{\mathrm{tr}})| \geqslant \varepsilon\} Y(t) p_j(t,\theta_{\mathrm{tr}}) \alpha_{\mathrm{tr}}(t)\, \mathrm{d}t \to_{\mathrm{p}} 0$$

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

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

which completes the proof of (ii). As for the convergence in probability in (iii), the first term in (22) converges to Σ by Condition (A) and the 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 (22) is $\mathrm{O_p}(n^{-1/2})$, 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 $\sqrt{n}(\widehat{\theta}_n-\theta_{\rm tr})=\Sigma^{-1}H_n(\tau)+\mathrm{o_p}(1)$ and

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

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

7.2. Proof of (16) in Theorem 2

By reasoning as in Theorem 1.4.2 of Ghosh and Ramamoorthi (2003), the convergence in (16) is implied by

$$\int_{\mathbb{R}^{2k-2}}|g_n(x)-\phi(x)\pi(\theta_{\rm tr})|\,\mathrm{d}x\to_{\rm p}0, \tag{23}$$

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 $\phi(x)=\exp\{-x^{\rm t}\Sigma x/2\}$. From this point on, we follow slightly different techniques as we borrow from the results on the conditional log-likelihood $\ell_n(\theta)$ previously obtained.

Proof of (23) goes along with the decomposition

$$\int_{\mathbb{R}^{2k-2}}|g_n(x)-\phi(x)\pi(\theta_{\rm tr})|\,\mathrm{d}x\leqslant\int_{|x|\leqslant K}|g_n(x)-\phi(x)\pi(\theta_{\rm tr})|\,\mathrm{d}x+\int_{K<|x|<\delta\sqrt{n}}g_n(x)\,\mathrm{d}x+\int_{|x|\geqslant\delta\sqrt{n}}g_n(x)\,\mathrm{d}x+\int_{|x|>K}\phi(x)\pi(\theta_{\rm tr})\,\mathrm{d}x.$$

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 $\varepsilon>0$, two positive constants K and δ such that $\Pr\{I_j>\varepsilon\}\to 0$ for $j=1,2,3$.

For I_1 we use the third-order Taylor expansion

$$\ell_n(\widehat{\theta}_n+x/\sqrt{n})-\ell_n(\widehat{\theta}_n)=-\tfrac{1}{2}x^{\rm t}\Sigma_n(\widehat{\theta}_n)x-\tfrac{1}{6}r_n(x,\theta_*), \tag{24}$$

for θ_* such that $|\theta_*-\widehat{\theta}_n|\leqslant|x|/\sqrt{n}$. The remainder $r_n(x,\theta_*)$ is defined as in (20) and $\Sigma_n(\widehat{\theta}_n)$ is defined accordingly to (19). Next, for $\phi_n(x)=\exp\{-x^{\rm t}\Sigma_n(\widehat{\theta}_n)x/2\}$,

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

The continuity of $V(t,\theta)$ with respect to θ and an application of Lenglart's inequality implies that $\Sigma(\widehat{\theta}_n)\to_{\rm p}\Sigma$, so that the first term goes to zero. As for the second term, we have

$$|g_n(x)-\phi_n(x)\pi(\theta_{\rm tr})|\leqslant\phi_n(x)\pi(\theta_{\rm tr})\left[(1+\eta_1)\sup_{|x|\leqslant K}|\exp\{-r_n(x,\theta_*)\}-1|+\eta_1\right],$$

where $\eta_1=\sup_{|x|\leqslant K}|\pi(\widehat{\theta}_n+x/\sqrt{n})/\pi(\theta_{\rm tr})-1|$. Note that $\eta_1\to_{\rm p}0$ because of $\pi(\widehat{\theta}_n+x/\sqrt{n})\to_{\rm p}\pi(\theta_{\rm tr})$ uniformly on $|x|\leqslant K$ and the positivity of $\pi(\theta_{\rm tr})$. Moreover, reasoning as in the proof of Theorem 1, one finds that, for any $K>0$ and $|x|\leqslant K$, $\sup_{|x|\leqslant K}|\exp\{-r_n(x,\theta_*)\}-1|\to_{\rm p}0$. Since $\int_{|x|\leqslant K}\phi_n(x)\,\mathrm{d}x$ is bounded in probability, we conclude that, for any $K>0$, $\Pr\{I_1>\varepsilon\}\to 0$.

Regarding I_2, we use the following bound for the third order term of the Taylor expansion in (24):

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

where the integral on the right-hand side is a quantity bounded in probability (see Conditions (A) and (B) and use Lenglart's inequality). Hence, there exists $M>0$ large enough such that $\Pr\{|r_n(x,\theta_*)|>\delta Mx^{\rm t}x\}\to 0$. 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 $x^{\rm t}B_nx\leqslant b_nx^{\rm t}x$; (ii) λ as the smallest eigenvalue of Σ, such that $x^{\rm t}\Sigma x\geqslant\lambda x^{\rm t}x$. Hence, we have $-\frac{1}{2}x^{\rm t}\Sigma_n(\widehat{\theta}_n)x\leqslant-\frac{1}{2}(\lambda-b_n)x^{\rm t}x$. Since $\lambda>0$ (Σ is positive definite) and $b_n\to_{\rm p}0$, for some ε_1 sufficiently small there exists $\delta=\delta(\lambda,M,\varepsilon_1)$ such that $\lambda+\delta M/3-\varepsilon_1>0$ and $\Pr\{\lambda-b_n-2\delta M<\varepsilon_1\}\to 0$. Fix now $\varepsilon>0$ and use expansion in (24) to get

$$\begin{aligned}\Pr\{I_2>\varepsilon\}\leqslant{}&\Pr\left\{\sup_{|x|\leqslant\delta\sqrt{n}}|\pi(\widehat{\theta}_n+x/\sqrt{n})/\pi(\theta_{\rm tr})|>\eta_2\right\}+\Pr\{|r_n(x,\theta_*)|>\delta Mx^{\rm t}x\}\\&+\Pr\{b_n>\lambda+\delta M/3-\varepsilon_1\}+\Pr\left\{\int_{K<|x|<\delta\sqrt{n}}\mathrm{e}^{-\varepsilon_1x^{\rm t}x/2}\,\mathrm{d}x>\frac{\varepsilon}{\eta_2\pi(\theta_{\rm tr})}\right\},\end{aligned}$$

where $\eta_2 = \sup_{|x| \leqslant 2\delta} |\pi(\theta_{\rm tr} + x)/\pi(\theta_{\rm tr})|$. The first term on the right-hand side goes to zero because $|\widehat{\theta}_n - \theta_{\rm tr}| \leqslant 2\delta$ 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 $\int_{|x|>K} \mathrm{e}^{-\varepsilon_1 x^{\rm t} x/2}\,\mathrm{d}x \leqslant \varepsilon/\eta_2 \pi(\theta_{\rm tr})$. Then, there exist $K, \delta > 0$ such that $\Pr\{I_2 > \varepsilon\} \to 0$.

As for I_3, consider that

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

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

$$n^{k-1} \sup_{|\theta - \theta_{\rm tr}| = \delta/2} \exp\{\ell_n(\theta) - \ell_n(\widehat{\theta}_n)\} \to_{\rm p} 0. \tag{25}$$

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

$$n^{k-1} \sup_{|\theta - \theta_{\rm tr}| = \delta/2} \exp\{\ell_n(\theta) - \ell_n(\widehat{\theta}_n)\} \leqslant n^{k-1} \sup_{|\theta - \theta_{\rm tr}| = \delta/2} \exp\{n[\mathrm{o}_{\rm p}(1) + d(\theta)]\}.$$

Hence, (25) holds because $n^{k-1}\mathrm{e}^{nd(\theta)} \to 0$. We conclude that, for any $\varepsilon, \delta > 0$, $\Pr\{I_3 > \varepsilon\} \to 0$, and the proof is complete.

8. Treatment for general p_j

As we have already pointed out, Eq. (2) corresponds to cause-specific conditional probabilities in the case of Gompertz-type CSHs, see Eq. (3). Other forms for p_j are equally plausible, e.g. one could start from a different parametric family of hazard rates. This suggests a natural way to generalize (2), a convenient form being

$$\begin{cases} p_j(t,\theta) = \mathrm{e}^{\lambda(t,\theta_j)} p_k(t,\theta), & j = 1,\ldots,k-1, \\ p_k(t,\theta) = 1 \Big/ \left\{ \sum_{h \leqslant k} \mathrm{e}^{\lambda(t,\theta_h)} \right\}. \end{cases} \tag{26}$$

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 (26) without serious efforts. First note that both the part regarding the overall hazard and the conditions for the prior distributions on A and θ remain the same. We need to replace Conditions (B) and (C) of Section 2 in order to make Theorems 1 and 2 hold, see Section 6. To this aim, we introduce some additional notation for the first and second derivatives of $\lambda(t,\theta_j)$ with respect to θ:

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

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

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

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

(B1′) $\int_0^\tau (\max_{j \leqslant k} |z_j(t,\theta)|)^4 \alpha_{\rm tr}(t)\,\mathrm{d}t < \infty$;
(B2′) $\int_0^\tau (\max_{j \leqslant k} |Z_j(t,\theta)_{h,i}|)^4 \alpha_{\rm tr}(t)\,\mathrm{d}t < \infty$ for any $h,\ i = 1,\ldots,2k-2$;
(B3′) $\{\ell^{(3)}_{n,hil}(\theta)\} = \mathrm{O}_{\rm p}(n),\ h,i,l, = 1,\ldots,2k-2$,

where, in (B3′), $\ell_n^{(3)}(\theta)$ denotes the array of the third derivatives of the log-likelihood of θ, now given by $\ell_n(\theta)=\sum_{j\leqslant k}\int_0^\tau[\lambda(t,\theta_j)+\log p_k(t,\theta)]\,\mathrm{d}N_j(t)$. Condition (C) is replaced by

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

The statement of Theorem 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 $H_n(\tau)\to_{\mathrm{d}}\mathrm{N}(0,\Sigma)$ whereas condition (B2′) guarantees that $\Sigma_n(\theta_{\mathrm{tr}})\to_{\mathrm{p}}\Sigma$. Condition (B3′) is needed for the remainder

$$r_n(x,\theta)=\frac{1}{\sqrt{n}}\sum_{h,i,l=1}^{2k-2}\frac{x_hx_ix_l}{6}\ell_{n,hil}^{(3)}(\theta)/n$$

to be asymptotically negligible for θ in a neighborhood of θ_{tr}. The asymptotic independence of $\widehat{A}_n$ and $\widehat{\theta}_n$ follows from arguments similar to those used in the proof of Theorem 1. We stress that, in general, one needs to check Condition (C′) case by case.

As for Theorem 2, asymptotic normality of the posterior distribution of θ can be proved under Conditions (A), (B1′), (B2′), (B3′) and (C′) with techniques similar to those used in Section 6.2. 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 $n^{-1}[\ell_n(\theta)-\ell_n(\theta_{\mathrm{tr}})]$ to a strictly concave function with maximum at θ_{tr} equal to zero. Condition (B3′) is involved in (i) and (ii), whereas (iii) is dealt via Conditions (B1′) and (B2′) in conjunction with Lenglart's inequality. Finally, Condition (C′) is the key tool for establishing (iv).

Remark 1. It is worth considering the subclass of models in (26) that arises when $\lambda(t,\theta_j)$ has the following form:

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

with g continuous. The same arguments used for model (2) apply 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 greater than one and $z_j(t)$ has $(1,g(t))^{\mathrm{t}}$ in the j-th block and zeros elsewhere. Condition (B) is replaced by assuming that $\int_0^\tau[g(t)^2+1]^3\alpha_{\mathrm{tr}}(t)\,\mathrm{d}t<\infty$, 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, overlap of the original failure times is sufficient when g is a monotone function, 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.

9. Concluding remarks

In the present paper we have investigated the BvM theorem for competing risks models under a semiparametric approach. We have established the asymptotic normality of the posterior distribution of differentiable functionals of model parameters in the theoretical framework of the multiplicative intensity model of Aalen. The semiparametric formulation induces a factorization in the likelihood that simplifies the derivation of the large-sample distribution of $(\widehat{A}_n,\widehat{\theta}_n)$. Since we get independence in the posterior distribution of A and θ, the BvM theorem for any differentiable functionals follows from the BvM theorem for A and θ and an application of the functional delta method. We have shown how a convenient formulation of the semiparametric model allows to exploit the good large-sample behavior of the beta process in order to derive BvM for functionals of interest without severe efforts.

Model (1) is interesting 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 a graphical device for describing the prevalence of risks, which is an important topic in competing risks models. Future work will focus on investigating the robustness of (2) and (26) to model misspecification and on including covariate effects in the cause-specific failure time distribution.

An alternative and fully nonparametric approach to the analysis of competing risks data might follow the methods set forth in Huang and Louis (1998). The competing risks problem is a special case where survival times are associated with mark variables which are not observed when the event is censored. Huang and Louis propose to estimate the joint distribution of survival times and mark variables by using a *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 then be of interest to develop Bayesian methods and study the asymptotic properties of the posterior distribution in this setting.

Acknowledgments

The authors are grateful to an Associate Editor and an anonymous referee for helpful comments. Moreover, Igor Prünster is gratefully acknowledged for some useful discussions that led to improvements in the manuscript.

References

Aalen, O., 1978. Nonparametric inference for a family of counting processes. Ann. Statist. 6, 701–726.
Albert, A., Anderson, J.A., 1984. On the existence of maximum likelihood estimates in logistic regression models. Biometrika 71, 1–10.
Andersen, P.K., Gill, R.D., 1982. Cox's regression model for counting processes: a large sample study. Ann. Statist. 10, 1100–1120.
Andersen, P.K., Borgan, Ø., Gill, R.D., 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., Hjort, N.L., 2007. Bayesian survival analysis in proportional hazard models with logistic relative risk. Scand. J. Statist. 34, 229–257.
Diaconis, P., 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., Karia, S.R., 2000. Analysis of competing risks by using Bayesian smoothing. Scand. J. Statist. 27, 605–617.
Ghosh, J.K., Ramamoorthi, R.V., 2003. Bayesian Nonparametric. Springer, Berlin.
Gilks, W.R., 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., 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., Pollard, D.B., 1994. Asymptotics for minimisers of convex processes. Statistical Research Report, Department of Mathematics, University of Oslo.
Huang, Y., Louis, T.A., 1998. Nonparametric estimation of the joint distribution of survival time and mark variables. Biometrika 85, 785–798.
Ibrahim, J.G., 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., Lee, J., 2004. A Bernstein–von Mises theorem in the nonparametric right-censoring model. Ann. Statist. 32, 1492–1512.
Larson, M.G., Dinse, G.E., 1985. A mixture model for the regression analysis of competing risks data. Appl. Statist. 34, 201–211.
Pepe, M.S., Mori, M., 1993. Kaplan–Meier, marginal and conditional probability curves in summarizing competing risks failure time data. Statist. Med. 12, 737–751.
Sato, K., 1999. Lévy Processes and Infinitely Divisible Distributions. Cambridge University Press, Cambridge.
Zhao, L.H., 2000. Bayesian aspects of some nonparametric problems. Ann. Statist. 28, 532–552.