Uncertainty quantification for Bayesian survival analysis

Ismaël Castillo[1] Stéphanie van der Pas[2]

[1]Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université

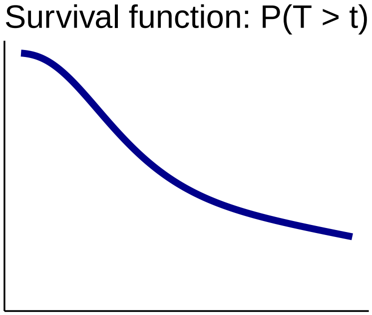
[2]Mathematical Institute, Leiden University

O'Bayes 2019, University of Warwick – June 30, 2019

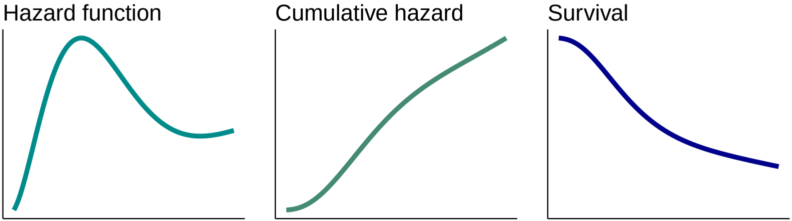
Survival model

T : event time; C : censoring time.
$Y = \min\{T, C\}$; $\delta = \mathbf{1}\{T \leq C\}$.
Independent right censoring.

We **observe** n independent pairs $(Y_1, \delta_1), \ldots, (Y_n, \delta_n)$.



Survival objects



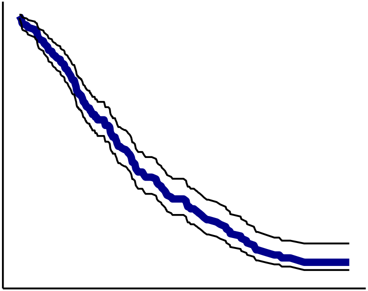
$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}.$$

$$\Lambda(t) = \int_0^t \lambda(u) du.$$

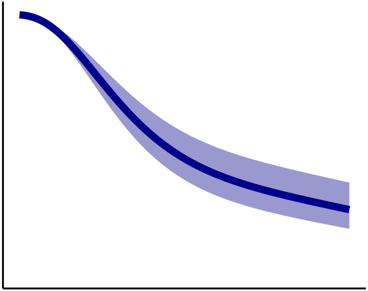
$$S(t) = e^{-\Lambda(t)}.$$

Goal

Pointwise intervals



Credible band



Bayesian survival analysis

- Kim and Lee (2004), Kim (2006), BvM for survival objects using neutral to the right processes.
- De Blasi, Peccati, Prünster (2009), CLTs for linear and quadratic functionals of the hazard, using kernel mixtures with respect to a completely random measure.
- De Blasi and Hjort (2009), BvM in competing risks setting, using a beta process prior.
- Castillo (2012), semiparametric BvM for the Cox model, using a Riemann-Liouville type process.
- Donnet, Rivoirard, Rousseau, Scricciolo (2017), L^1-concentration for priors where the normalized hazard and its integral are specified independently.

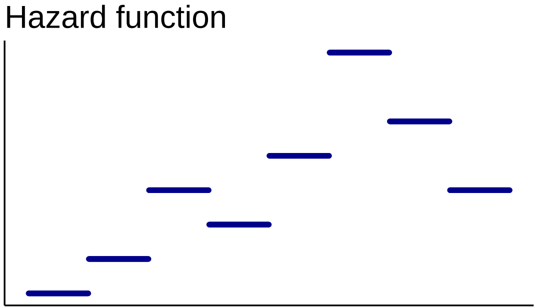
Our approach

Our setting is non-conjugate and the techniques potentially apply to many priors.

Key result: establish convergence at the minimax rate in $\|.\|_\infty$-norm.
Using tools from Castillo (2012), Castillo (2014), Castillo & Rousseau (2015).

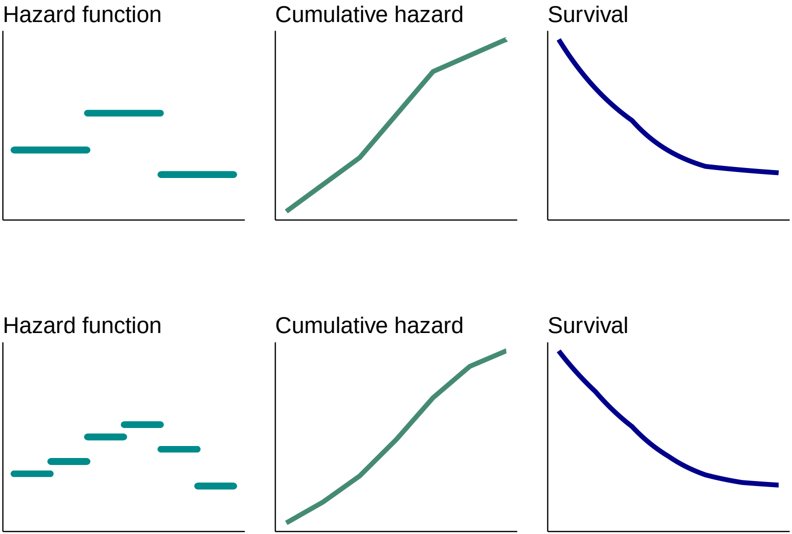
Next, obtain nonparametric BvM in appropriately weighted multiscale space, and use continuous mapping and the functional delta method.
As in Castillo & Nickl (2014), Gill (1989).

The piecewise exponential model



E.g. Gamerman (1991, 1994), Ibrahim, Chen and Sinha (2013), Kalbfleisch (1978), Arjas and Gasbarra (1994), Nieto-Barajas and Walker (2002).

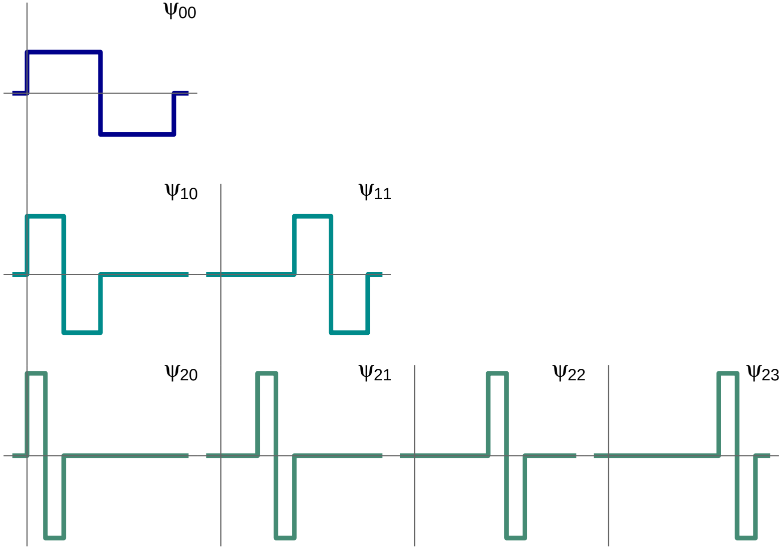
Rough hazard, smooth survival



Haar wavelets

$\psi_{lk}(t) = 2^{l/2}\psi(2^l t - k), 0 \le k < 2^l.$

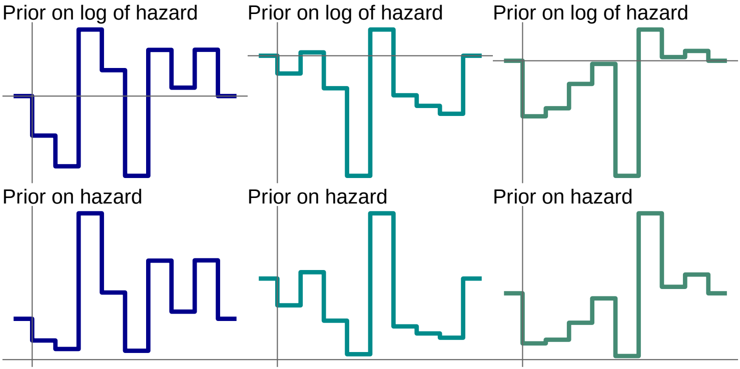
$\psi(t) = \mathbf{1}\{t \in (0, 1/2]\} - \mathbf{1}\{t \in (1/2, 1]\};$



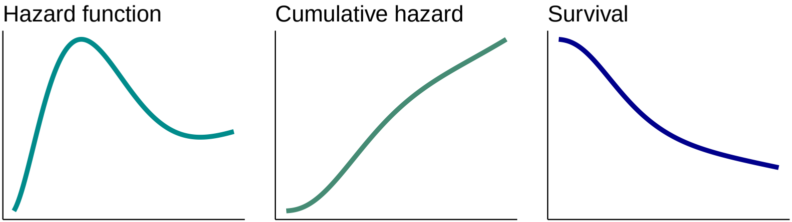
Haar-histogram priors

Defined on $r = \log \lambda$, with given cut-off L_n:

$$r = \sum_{l=-1}^{L_n} \sum_{k=0}^{2^l - 1} Z_{lk} \psi_{lk}, \quad Z_{lk} \text{ independent r.v.}$$



Assumptions



- There exists $\tau > 0$ such that $S_0(\tau) > 0$ and $\mathbb{P}_{\lambda_0}(C \geq \tau) > 0$. We take $\tau = 1$.
- There exists $\rho > 0$ such that $\lambda_0(t) \geq \rho$ for all $t \in [0, \tau]$.
- There exist $c_1, c_2 > 0$ such that $\Lambda_0(\tau) < c_1$, $\|\lambda_0\|_\infty < c_2$.
- C admits a continuous density g with respect to the Lebesgue measure on $[0, \tau)$.
- There exists $\rho' > 0$ such that $g(t) \geq \rho'$ for all $t \in [0, \tau)$.

Smoothness

For $\beta > 0$, $M > 0$ and a set $\psi = \{\psi_{lk}\}$:

$$\mathcal{C}_{\psi}^{\beta,M} = \left\{ \lambda \in C([0,1]) : \sup_{l \geq 0} \max_{0 \leq k < 2^l} 2^{l(\beta+1/2)} |\langle \psi_{lk}, \lambda \rangle_2| \leq M \right\}.$$

For the standard Hölder class $\mathcal{H}(\beta, M)$ and any $B > 0$:

- $\mathcal{H}(\beta, M) \subset \mathcal{C}_{\psi}^{\beta,M}$ for $\beta \leq 1$ when $\{\psi_{lk}\}$ are the Haar wavelets.
- $\mathcal{H}(\beta, M) \subset \mathcal{C}_{\psi}^{\beta,M}$ for all $\beta \leq B$ if $\{\psi_{lk}\}$ is smooth enough.

$\|.\|_\infty$-rate result for Haar-histogram priors

Theorem. Let $X^{(n)} = (Y_1, \delta_1), \ldots, (Y_n, \delta_n)$ be the observations. Suppose $\lambda_0 \in \mathcal{C}_\psi^{\beta,M}$ with $\beta > 1/2$, some $M > 0$, and ψ the Haar wavelets. Set $\alpha = \min\{\beta, 1\}$. Let the prior on the hazard be the Gaussian Haar-histogram:

$$r = \sum_{l=-1}^{L_n} \sum_{k=0}^{2^l-1} Z_{lk}\psi_{lk}, \quad Z_{lk} \sim \mathcal{N}(0, \sigma_l^2),$$

where $\sigma_l = 2^{-l(1/2+\alpha)}$ and $L_n = \lfloor \log_2\{(n/\log n)^{1/(2\alpha+1)}\}\rfloor$.

Then, with $\varepsilon^*_{n,\alpha} = (\frac{\log n}{n})^{\alpha/(2\alpha+1)}$, there exists $M > 0$ such that:

$$\mathbb{E}^n_{\lambda_0} \int \|\lambda - \lambda_0\|_\infty d\Pi(\lambda \mid X^{(n)}) \leq M\varepsilon^*_{n,\alpha}.$$

Major steps in the proof

- Prove L_1-concentration;
- Establish a BvM result for linear functionals $\langle b, \lambda \rangle_2 = \int_0^\tau b(u)\lambda(u)du$;
  where b is a bounded function.
- Relate $\|.\|_\infty$ to $\{\langle \psi_{lk}, \lambda \rangle_2\}$ and use tightness at rate $n^{-1/2}$ (from previous step) on this collection.

From Hellinger to L_1 concentration

Lemma

On a set where $h^2(p_\lambda, p_{\lambda_0}) \lesssim \varepsilon_n^2$, we have: $\|\lambda - \lambda_0\|_1^2 \lesssim \varepsilon_n^2$.

The squared Hellinger distance $h^2(p_\lambda, p_{\lambda_0}) = h^2(p_{\lambda,g}, p_{\lambda_0,g})$ is given by

$$
\begin{aligned}
h^2(p_\lambda, p_{\lambda_0}) &= \int_0^\tau \left[\sqrt{gS} - \sqrt{gS_0}\right]^2 (y)dy \\
&\quad + \int_0^\tau \bar{G}(y) \left[\sqrt{\lambda S} - \sqrt{\lambda_0 S_0}\right]^2 (y)dy \\
&\quad + \bar{G}(\tau) \left[\sqrt{S} - \sqrt{S_0}\right]^2 (\tau).
\end{aligned}
$$

BvM for linear functionals of the hazard

Write $\psi(\lambda) = \langle b, \lambda \rangle_2 = \int_0^\tau b(u)\lambda(u)du$.

We intend to show for all $t \in \mathbb{R}$,

$$\mathbb{E}\left[e^{t\sqrt{n}[\psi(\lambda)-\widehat{\psi}]} \mid X^n, A_n\right] \overset{P_{\lambda_0}}{\to} e^{\frac{t^2}{2}\|\tilde{\psi}_0\|_L^2},$$

for some 'efficient estimator' $\widehat{\psi}$, with $\tilde{\psi}_0 = b/M_0$, and $M_0(u) = P_0(Y \geq u)$.

Then a BvM result follows from the results in Section 1 of Castillo and Rousseau (2015), Supplement.

Introducing an efficient centering

$$\|\lambda - \lambda_0\|_\infty \leq \|\lambda - \lambda^*_{L_n}\|_\infty + \|\lambda^*_{L_n} - \lambda_{0,L_n}\|_\infty + \|\lambda_{0,L_n} - \lambda_0\|_\infty,$$

where

- $\lambda^*_{L_n}$ is an 'efficient centering';
- λ_{0,L_n} is the L^2-projection of λ_0 on Vect$\{\psi_{lk}, l \leq L_n\}$.

$\|.\|_\infty$ and wavelet coefficients [Castillo (2014)]

$$\|\lambda - \lambda^*_{L_n}\|_\infty \leq \sum_{l=0}^{L_n} \max_{0\leq k<2^l} |\langle \lambda - \lambda^*_{L_n}, \psi_{lk}\rangle| \Big\| \sum_{0\leq k<2^l} |\psi_{lk}| \Big\|_\infty$$

$$\leq \frac{1}{\sqrt{n}} \sum_{l=0}^{L_n} 2^{l/2}\sqrt{n} \max_{0\leq k<2^l} |\langle \lambda - \lambda^*_{L_n}, \psi_{lk}\rangle|.$$

For any $t > 0$:

$$\mathbb{E}_{\lambda_0}\mathbb{E}^{\Pi_n} \max_{0\leq k<2^l} t\sqrt{n}|\langle \lambda - \lambda^*_{L_n}, \psi_{lk}\rangle_2|$$

$$\leq \log\left(\sum_{k=0}^{2^l-1} \mathbb{E}_{\lambda_0}\mathbb{E}^{\Pi_n}\left[e^{t\sqrt{n}\langle \lambda-\lambda^*_{L_n},\psi_{lk}\rangle_2} + e^{-t\sqrt{n}\langle \lambda-\lambda^*_{L_n},\psi_{lk}\rangle_2}\right]\right)$$

BvM for functionals

On sets $A_n = \{\lambda : \|\lambda - \lambda_0\|_1 \leq \varepsilon_n\}$, there exists a $C > 0$ independent of l, k such that, for any $t > 0$:

$$\mathbb{E}\left[e^{t\sqrt{n}\langle\lambda - \lambda^*_{L_n}, \psi_{lk}\rangle_2} \mid X^n, A_n\right] \leq e^{\frac{t^2}{2}C}(1 + o_p(1)).$$

We intend to show for all $t \in \mathbb{R}$,

$$\mathbb{E}\left[e^{t\sqrt{n}[\psi(\lambda) - \widehat{\psi}]} \mid X^n, A_n\right] \overset{P_{\lambda_0}}{\to} e^{\frac{t^2}{2}\|\tilde{\psi}_0\|_L^2},$$

for some 'efficient estimator' $\widehat{\psi}$, with $\tilde{\psi}_0 = b/M_0$, and $M_0(u) = P_0(Y \geq u)$.

Nonparametric BvM [Castillo and Nickl (2013, 2014)]

A ‘nonparametric’ Bernstein-von Mises can be stated as, for some centering T_n:

$$\mathcal{L}(\sqrt{n}(\lambda - T_n) \mid X^{(n)}) \stackrel{??}{\to} (\square)$$

Ingredients:

- limiting distribution $(\square)$ and ‘efficient estimator’ T_n;
- rate of convergence $\sqrt{n}$;
- sense of the convergence $\stackrel{??}{\to}$.

Multiscale approach: encode λ by its coefficients onto a wavelet basis

$$\{\psi_{lk}\}_{l\in\mathbb{N},\, 0\le k<2^l}.$$

Nonparametric multiscale BvM for histogram priors

Theorem. Let $\lambda_0 \in \mathcal{C}^\beta[0,1]$, for $\beta > 1/2$ and work under the same conditions as the sup-norm result for Π. For such Π,

$$\beta_{\mathcal{M}_0(w)}(\Pi(\cdot \mid X) \circ \tau_{T_n}^{-1}, \mathbb{G}_{\Lambda_0}) \to 0,$$

where β_S denotes the *bounded-Lipschitz metric* for weak convergence on metric space S.
Let w_l be a sequence such that $w_l/\sqrt{l} \uparrow \infty$.
$\mathcal{M}_0 := \mathcal{M}_0(w) = \left\{ f = \{\langle f, \psi_{lk}\rangle\}, \ \lim_{l\to\infty} \max_k \frac{|\langle f, \psi_{lk}\rangle|}{w_l} = 0 \right\}$

For the centering, with $\widehat{M}_0(y) = \frac{1}{n}\sum_{j=1}^n \mathbf{1}\{Y_j \geq y\}$:

$$T_n = \sum_{l \leq L_n} \sum_{k=0}^{2^l-1} \hat{\lambda}_{lk}\psi_{lk}, \quad \hat{\lambda}_{lk} = \frac{1}{n}\sum_{i=1}^n \delta_i \frac{\psi_{lk}}{\widehat{M}_0}(Y_i),$$

with cut-off L_n as before. The limiting distribution $\mathbb{G}_{\Lambda_0}$ is a white noise process defined as $E[\mathbb{G}_{\Lambda_0}(f)\mathbb{G}_{\Lambda_0}(g)] = \int fg\frac{\lambda_0}{M_0}$.

Using continuous mapping [Castillo and Nickl (2014)]

- We obtain a BvM result for the hazard in the multiscale space $\mathcal{M}_0$.
- By the continuous mapping theorem, from the BvM in $\mathcal{M}_0$ one can deduce limiting shape results for *continuous functionals*.
- It turns out that 'integration' is a continuous map.

Nonparametric BvM for cumulative hazard Λ

Let $\hat{\Lambda}_n$ be an efficient estimator of Λ.

Let $G_{\Lambda_0}(t) = W(U_0(t))$ with W Brownian motion and $U_0(t) = \int_0^t \frac{\lambda_0}{M_0}(u)du$, where $M_0(u) = \mathbb{E}_{\lambda_0}\mathbf{1}\{Y \geq u\}$.
Cf. Kim and Lee (2004).

Theorem. Let $\lambda_0 \in \mathcal{C}^\beta[0,1]$, for $\beta > 1/2$ and work under the same conditions as the sup-norm result for Π.

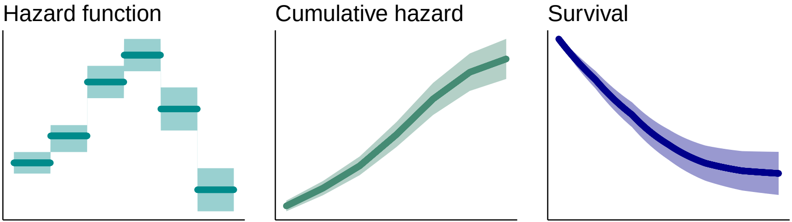
$$\beta_{L^\infty[0,1]}\Big(\mathcal{L}(\sqrt{n}(\Lambda - \hat{\Lambda}_n)|\, X^{(n)})\,,\, \mathcal{L}(G_{\Lambda_0})\Big) \to^{P_0} 0,$$

as well as

$$\beta_{\mathbb{R}}\Big(\mathcal{L}(\sqrt{n}\|\Lambda - \hat{\Lambda}_n\|_\infty|\, X^{(n)})\,,\; \mathcal{L}(\|G_{\Lambda_0}\|_\infty)\Big) \to^{P_0} 0.$$

Credible bands

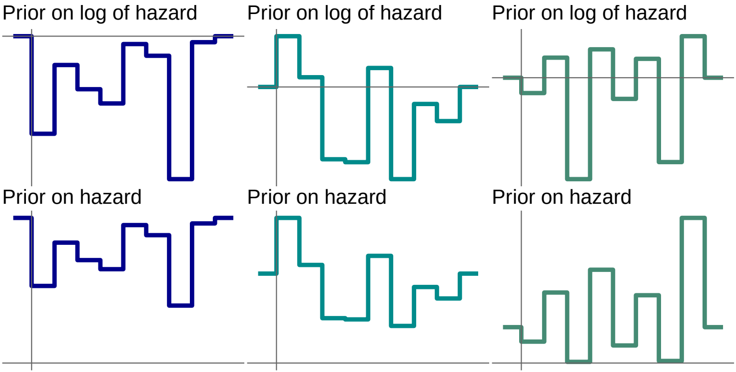
Consequence of previous theorem: credible bands for Λ are asymptotically optimal confidence bands.



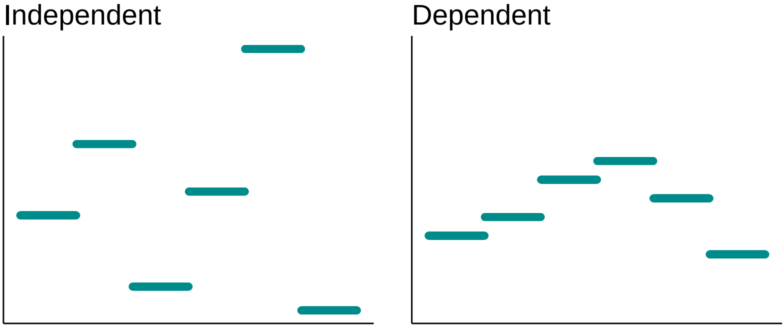
Dyadic histogram priors

With $I_k^J = [k2^{-J}, (k+1)2^{-J})$, $k = 1, \ldots, 2^J$:

$$r = \sum_{k=1}^{2^J} Z_k I_k^J, \quad Z_k \text{ independent r.v.}$$



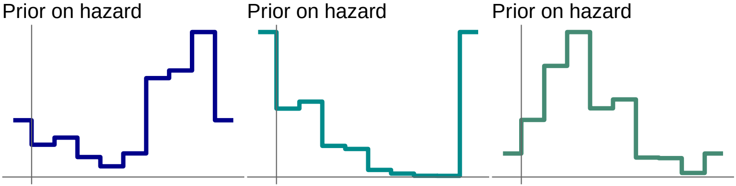
”Autoregressive” priors



Autoregressive histogram priors

Building on the autoregressive idea in Arjas and Gasbarra (1994), we construct dependent histograms such that, with λ_k the value on interval I_k^J on the hazard-scale:

$$\mathbb{E}[\lambda_k \mid \lambda_{k-1}, \ldots, \lambda_1] = \lambda_{k-1}.$$
$$\mathrm{Var}(\lambda_k \mid \lambda_{k-1}, \ldots, \lambda_1) = \sigma^2(\lambda_{k-1})^2.$$



Data example: dependent Gamma prior

Assume K intervals. Augment the data $\{(y_i, \delta_i)\}_{i=1}^n$:

- $y_i \to (y_{1i}, y_{2i}, \ldots, y_{Ki})$
  where y_{ki} is the total time individual i was observed in interval k.

- $\delta_i \to (\delta_{1i}, \delta_{2i}, \ldots, \delta_{Ki})$
  where δ_{ki} is equal to 1 if individual i has the event in interval k, and 0 otherwise.

$\vec{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_K)$, where λ_k is the hazard in interval k.

Then the likelihood can be written as [Holford (1980), Laird and Olivier (1981)]:

$$\ell(\vec{\lambda}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \lambda_k^{\delta_{ki}} e^{-\lambda_k y_{ki}} = \prod_{k=1}^{K} \lambda_k^{d_k} e^{-\lambda_k t_k},$$

where $d_k = \sum_{i=1}^n \delta_{ki}$ and $t_k = \sum_{i=1}^n y_{ki}$.

Data example: dependent Gamma prior

Following Arjas and Gasbarra (1994):

$$\lambda_1 \sim \text{Gamma}(\alpha_0, \beta_0)$$
$$\lambda_k \mid \lambda_1, \ldots, \lambda_{k-1} \sim \text{Gamma}(\alpha, \alpha/\lambda_{k-1}), \quad k = 2, \ldots, K.$$

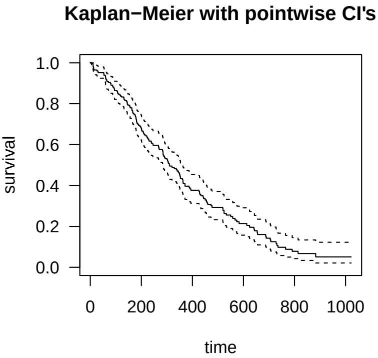
Then

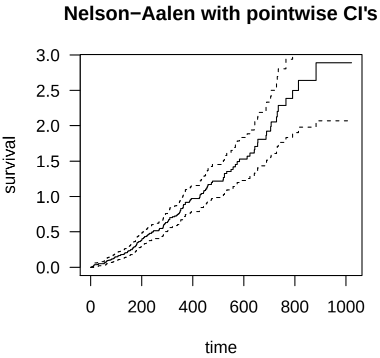
$$\mathbb{E}[\lambda_k \mid \lambda_1, \ldots, \lambda_{k-1}] = \lambda_{k-1}$$
$$\text{Var}(\lambda_k \mid \lambda_1, \ldots, \lambda_{k-1}) = \lambda_{k-1}^2/\alpha.$$

Data example: dependent Gamma prior

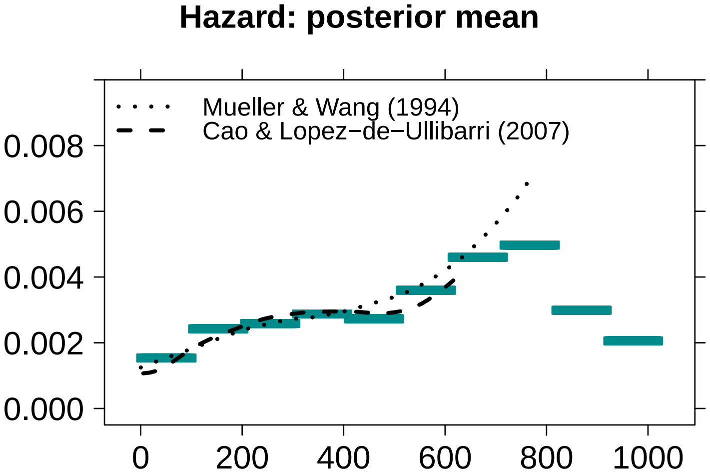
NCCTG Lung Cancer data, available in the R survival package.
165 deaths, 63 censored observations.

$K = 10$, $\alpha_0 = 1, \beta_0 = 1, \alpha = 2$.



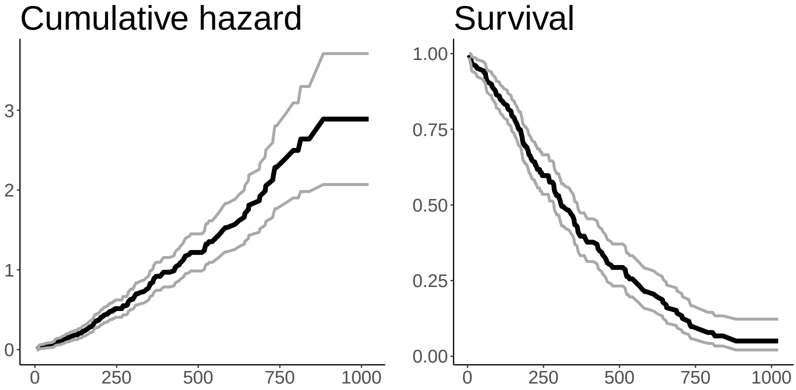


Data example: hazard

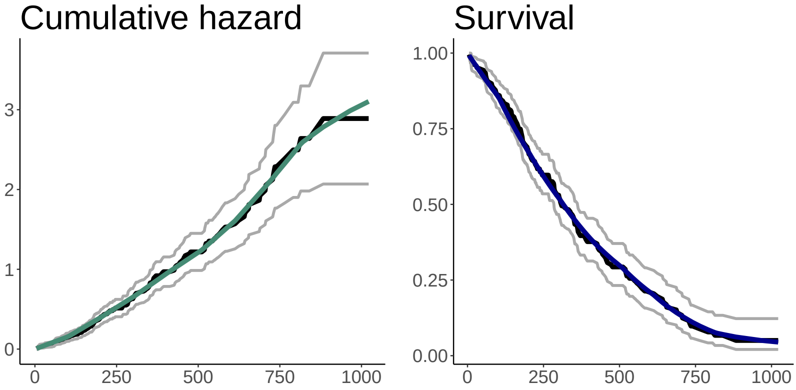


bin	1	2	3	4	5	6	7	8	9	10
events	31	44	30	21	12	11	9	6	1	0

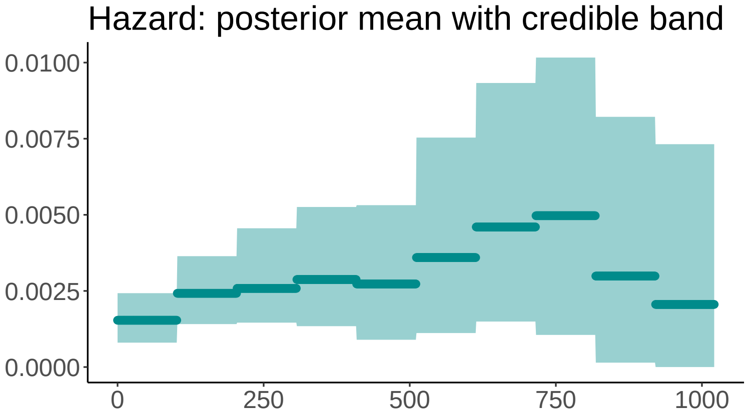
Data example: cumulative hazard and survival



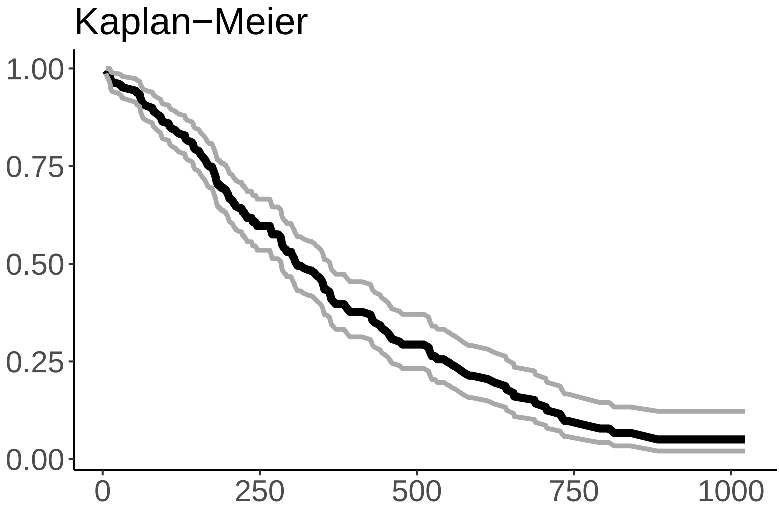
Data example: cumulative hazard and survival



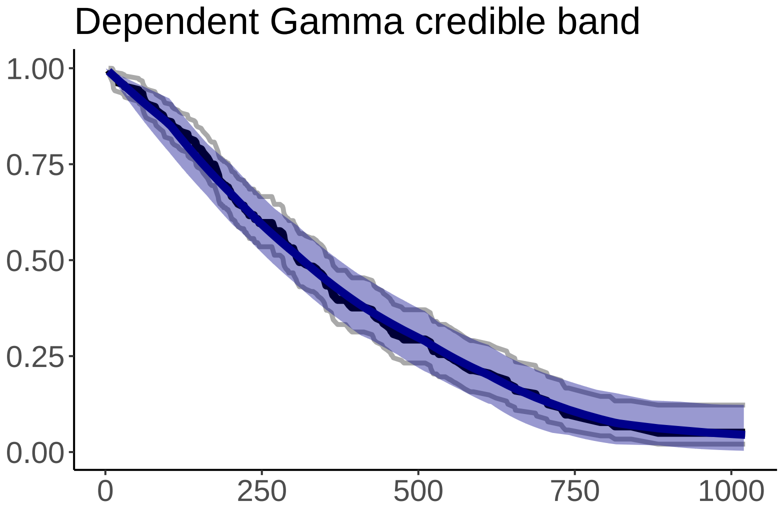
Data example: credible band for the hazard



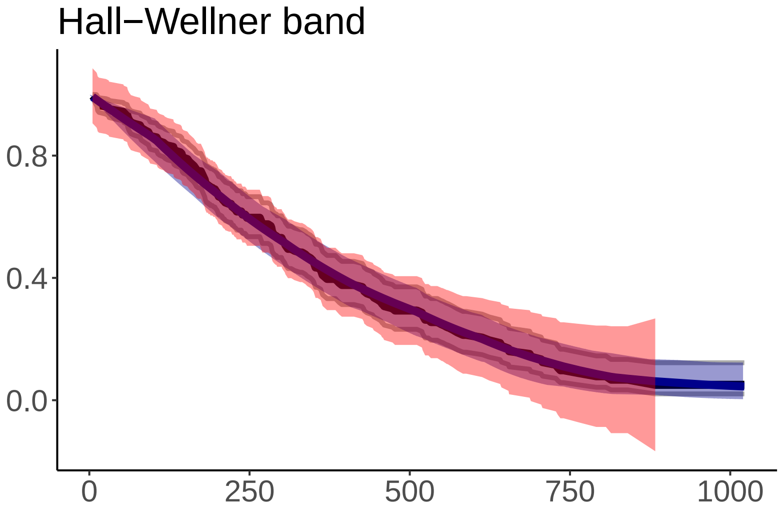
Data example: credible band for the survival



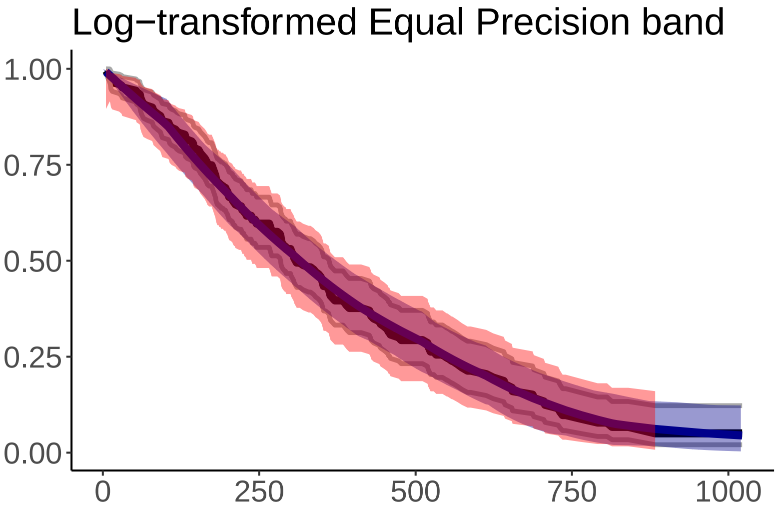
Data example: credible band for the survival



Data example: credible band for the survival



Data example: credible band for the survival



Data example: dependent lognormal prior

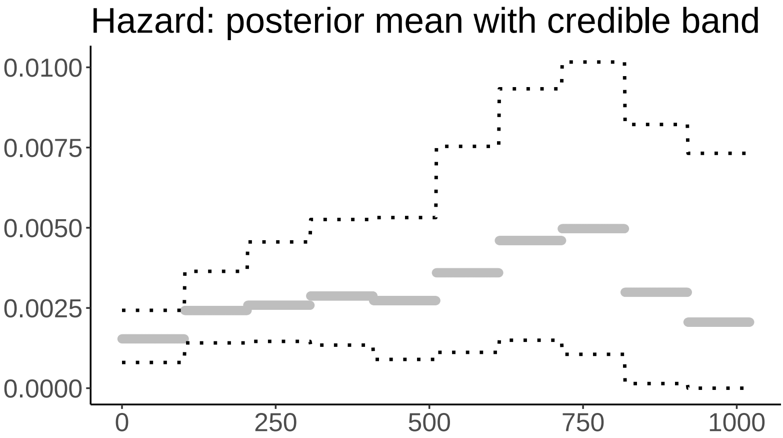
Write $r_k = \log \lambda_k$.

Prior:

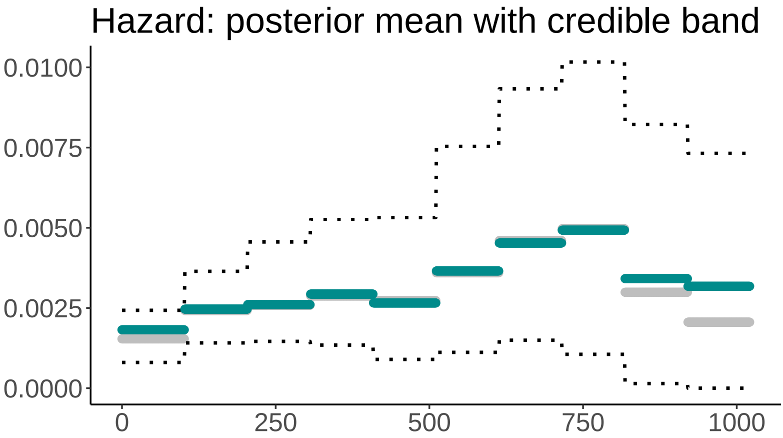
$$
\begin{gathered}
r_1 \sim \mathcal{N}(\mu_0, \sigma_0^2) \\
r_k \mid r_1, \ldots, r_{k-1} \sim \mathcal{N}(r_{k-1} - \tfrac{1}{2}\log(1+\sigma^2), \log(1+\sigma^2)) \\
k = 2, \ldots, K.
\end{gathered}
$$

In the data analysis, we set $\sigma = 1, \mu_0 = 0, \sigma_0^2 = \log(1+\sigma^2)$.

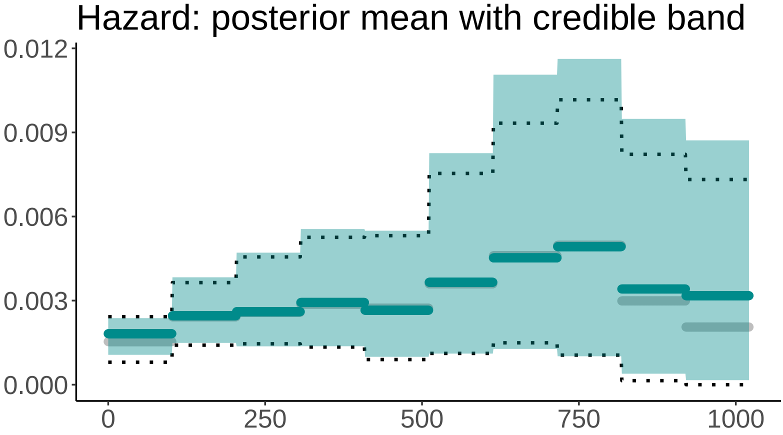
Gamma (grey/black) vs Lognormal (green)



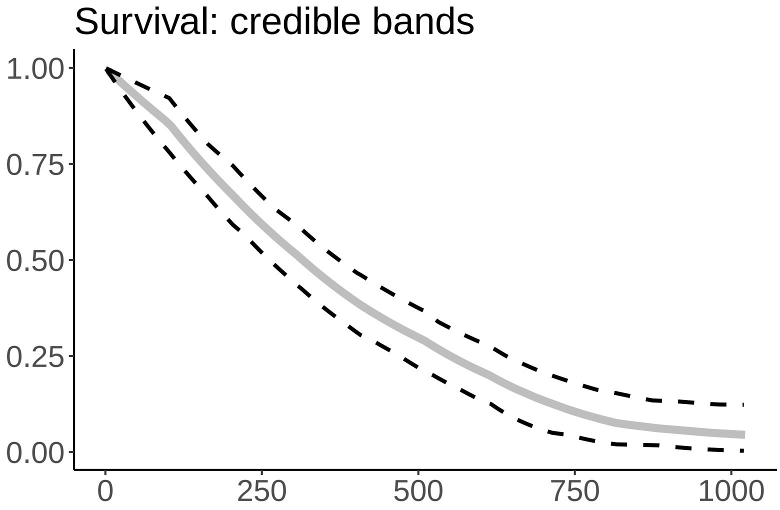
Gamma (grey/black) vs Lognormal (green)



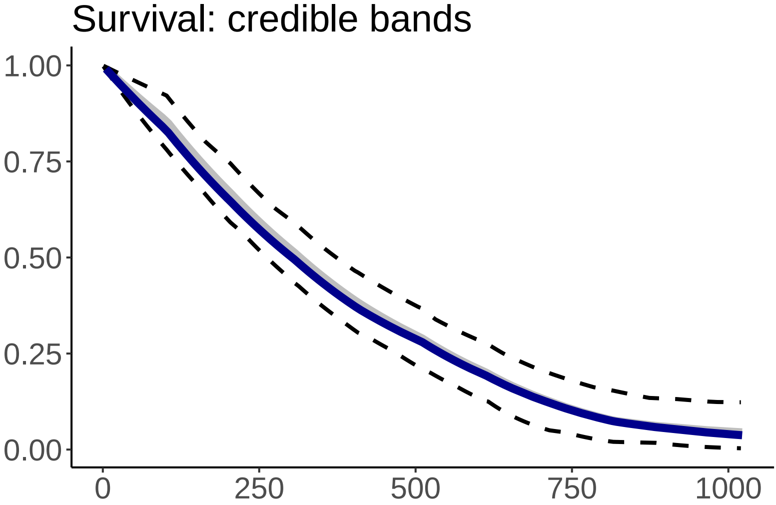
Gamma (grey/black) vs Lognormal (green)



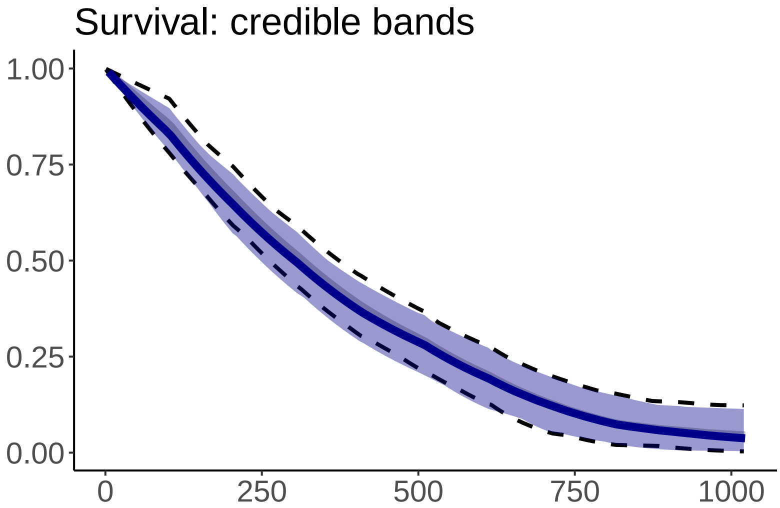
Gamma (grey/black) vs Lognormal (green)



Gamma (grey/black) vs Lognormal (green)

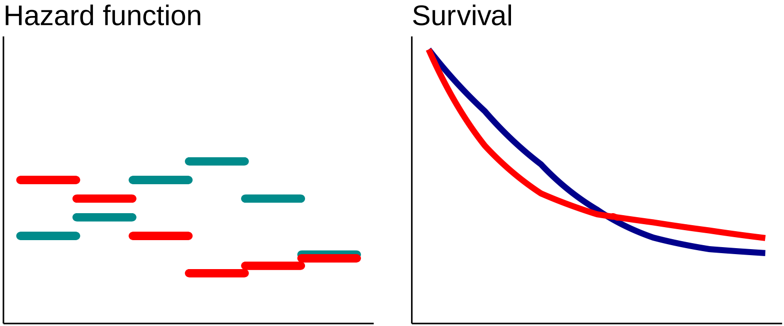


Gamma (grey/black) vs Lognormal (green)



Outlook: non-proportional hazards

E.g. Laird and Olivier (1981).



References

Arjas E and Gasbarra D (1994). Nonparametric Bayesiana inference from right censored survival data, using the Gibbs sampler. *Statistica Sinica*, 505–24.

Castillo I (2012). A semiparametric Bernstein-von Mises theorem for Gaussian process priors. *Probability Theory and Related Fields* **152**, 53–99.

Castillo I (2014). On Bayesian supremum norm contraction rates. *The Annals of Statistics* **42**, 2058–91.

Castillo I and Nickl R (2014). On the Bernstein-von Mises phenomenon for nonparametric Bayes procedures. *The Annals of Statistics* **42**, 1941–69.

Castillo I and Rousseau J (2015). A Bernstein-von Mises theorem for smooth functionals in semiparametric models. *The Annals of Statistics* **43**, 2353–83.

De Blasi P and Hjort N (2009). The Bernstein-von Mises theorem in semiparametric competing risks models. *Journal of Statistical Planning and Inference* **139**, 2316–28.

De Blasi P, Peccati G and Prünster I (2009). Asymptotics for posterior hazards. *The Annals of Statistics* **37**, 1906–45.

Donnet S, Rivoirard V, Rousseau J and Scricciolo C (2017). Posterior concentration rates for counting processes with Aalen multiplicative intensities. *Bayesian Analysis* **12**, 53–87.

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

Gamerman D (1994). Bayes estimation of the piece-wise exponential distribution. *IEEE Transactions on Reliability* **43**, 128–31.

Gill R (1989). Non- and semi-parametric maximum likelihood estimators and the Von Mises method, Part 1. *Scandinavian Journal of Statistics* **16**, 97–128.

Ibrahim JG, Chen MH, Sinha D (2013). *Bayesian survival analysis.* Springer Science & Business Media.

Kalbfleisch JD (1978). Non-parametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 241–21.

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

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

Nieto-Barajas LE and Walker SG (2002). Markova beta and gamma processes for modelling hazard rates. *Scandinavian Journal of Statistics* **29**, 413–24.