Tree-based Bayesian Mixture Model for Competing Risks

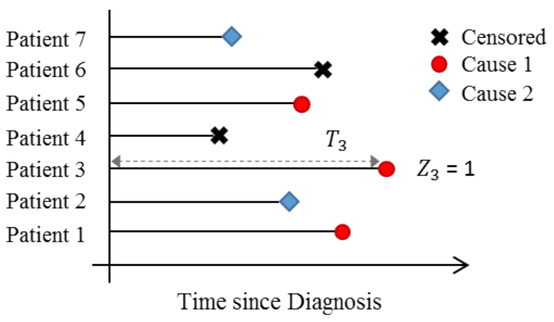
ALEXIS BELLOT* AND MIHAELA VAN DER SCHAAR*,†, † ALAN TURING INSTITUTE, * UNIVERSITY OF OXFORD

COMPETING RISKS

• **Competing Risks Analysis:** In survival analyses, competing risks hinder the observation of an event or modify the chance that it occurs. The fundamental problem is that at most one of the event times is observed in actual data, that leads to the need for specialized methods.

• **Example 1:** Receiving a **renal transplant** competes with **death on dialysis** for **kidney patients**.

• **Example 2:** Prognosis for **multi-morbidities** (e.g. cancer and cardiovascular diseases).



MIXTURE OF SURVIVAL DISTRIBUTIONS

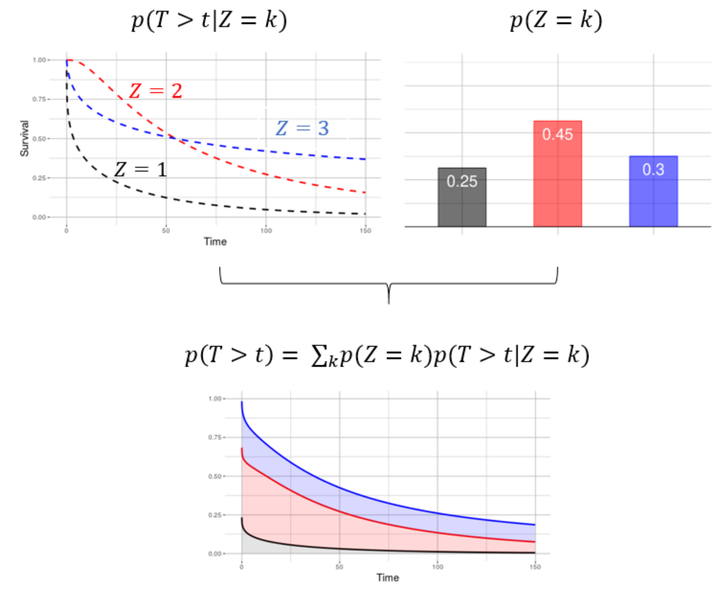
• To find a shared representation for the subject's survival with respect to multiple correlated co-morbidities, survival is represented as a mixture cause-specific trajectories with latent weights.

• **Bayesian approach to learning:** We update the posterior distribution of the event occurrence given observational data $\mathcal{D}$ and prior model $\boldsymbol{\theta}$. Cause-specific distributions are updated jointly to exploit relationships between co-morbidities.

▷ Uncertainty is represented directly on the survival curves.

• **Nonparametric association rules:** Variable interactions do not assume any predefined parametric form since these are generated with a nonparametric function $g(.)$.

▷ Overcomes linearity of variable interactions and proportional hazards of Fine-Gray and Cox models.



• **Important for** *optimal treatment planning* **and discovering** *shared biology*!

HIERARCHICAL MIXTURE MODEL (HBM)

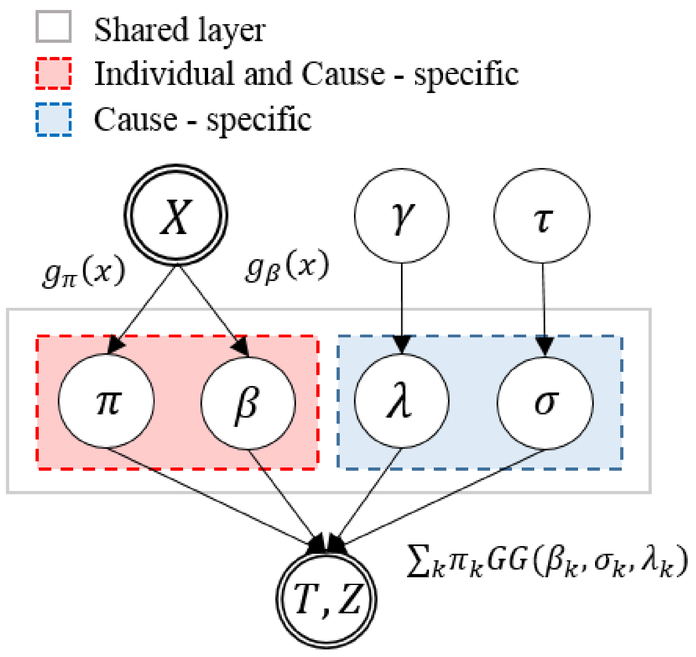
• **Generalised Gamma Survival Time Model:** Patient trajectories are generated as follows,

$$T_i|Z_i = k \sim \mathcal{GG}(\beta_{ik}, \sigma_i, \lambda_i)$$
$$Z_i \sim Cat(\pi_{i1}, ..., \pi_{iK})$$

◦ **The Mixture Regression Model:** Latent variables $\boldsymbol{\beta}_i = (\beta_{i1}, ..., \beta_{iK})$ and $\boldsymbol{\pi}_i = (\pi_{i1}, ..., \pi_{iK})$ propagate the influence of covariates through a multivariate random forest $g.(\boldsymbol{x}_i)$.

$$\boldsymbol{\beta}_i|\boldsymbol{x}_i \sim g_\beta(\boldsymbol{x}_i) + \epsilon_{\beta,i}, \quad \epsilon_{\beta,i} \sim \mathcal{N}(0, \alpha_\beta^2)$$
$$\boldsymbol{\pi}_i|\boldsymbol{x}_i \sim l(g_\pi(\boldsymbol{x}_i) + \epsilon_{\pi,i}), \quad \epsilon_{\pi,i} \sim \mathcal{N}(0, \alpha_\pi^2)$$

In every iteration prior means of $\boldsymbol{\beta}_i$ and $\boldsymbol{\pi}_i$ are updated with a sampled multi-output function g to inform the posterior.



• **Posterior Variable Importance:** The variable configurations that are strongly supported by the data will appear in most of the MCMC samples of g, while others with less evidence will appear less often.

▷ Accounts for the uncertainty in variable interactions.
▷ Distinguishes between influence on absolute risk of a specific end-point (e.g. probability of death due to CVD as opposed to Cancer) and influence on median survival for a specific cause.
▷ Enables Bayesian variable selection by controlling for a desired global false discovery rate bound.

◦ **Survival Shape Model:** The shape of the hazard (instantaneous risk of event) for each cause-specific survival distribution may exhibit different forms in *heterogeneous* populations. Allow for this flexibility by setting:

$$\sigma_i \sim \sum_{j=1}^{K} \mathbf{1}\{i \in \mathcal{C}_j\}\tau_j \text{ with } \tau_j \sim \mathcal{G}(\eta_0, \eta_1) \text{ and } \lambda_i \sim \sum_{j=1}^{K} \mathbf{1}\{i \in \mathcal{C}_j\}\gamma_j \text{ with } \gamma_j \sim \mathcal{N}(\gamma_0, \gamma_1).$$

POSTERIOR INFERENCE BY SAMPLING

• **Metropolis within Gibbs MCMC:** We sample parameters sequentially from a Markov Chain whose distribution converges to the true posterior.

▷ The contribution to the likelihood of an individual with end-point t_i of cause k is $\pi_{ik} f_k(t_i; \boldsymbol{\theta})$ while a censored observation at t_i contributes $S(t_i; \boldsymbol{\theta}) = 1 - F(t_i; \boldsymbol{\theta})$.

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_i \sum_k (\pi_{ik} f_k(t_i; \boldsymbol{\theta}))^{\mathbf{1}\{z_i=k\}} S(t_i; \boldsymbol{\theta})^{\mathbf{1}\{z_i=0\}}$$

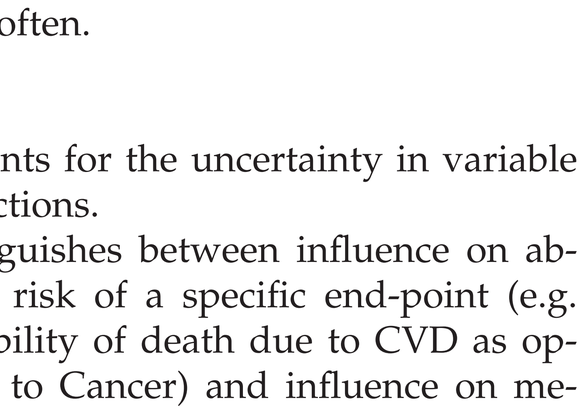
▷ The prior is the product of individual latent parameter prior distributions.
▷ Adaptive proposals are used to speed up convergence.

• **Prior Updates:** g_π and g_β are sampled to update the prior means of $\boldsymbol{\pi}$ and $\boldsymbol{\beta}$. A patient's covariates indirectly influence posterior parameter estimates through informed priors.

EXPERIMENTS

• **SEER dataset:** 72,000 patients diagnosed with breast cancer. 13.64% deaths due to breast cancer, 4.62% due to Cardiovascular diseases (CVD), and 7.3% due to other causes.

• **Benchmarks:** Cox, Fine-Gray model (FG), and Competing risks survival forest (CRF).



Agorithms	Cancer	CVD	Other
Cox	0.63 ± 0.02	0.60 ± 0.08	0.64 ± 0.07
FG	0.62 ± 0.01	0.59 ± 0.06	0.63 ± 0.06
CRF	0.69 ± 0.04	0.65 ± 0.09	0.68 ± 0.04
HBM	0.72 ± 0.02	0.64 ± 0.07	0.71 ± 0.08

Table 1: $C-$index (Higher better) on SEER with standard deviations.