1 Data Generation

Define p = 10 baseline covariates.

1.1 Risk predictors

We first generated a multivariate Gaussian p-vector $\tilde{\mathbf{X}}$ with exchangeable correlation, as follows:

$$\tilde{\mathbf{X}} \sim N\left(\mathbf{0}_p, (1-\rho)\mathbb{I}_p + \rho\right), \rho = 0.1,$$

where $\mathbf{0}_p$ is the p-dimensional zero vector, and $\mathbb{I}_p$ is the p-dimensional identity matrix. Next, we transformed $\tilde{X}_1, \ldots, \tilde{X}_{10}$ into data types commonly found in EHR datasets:

- Binary variables $X_j = \mathrm{I}\left(\tilde{X}_j > 0.3\right), j = 1, 2, 3, 4;$
- Count variables $X_j \sim \text{Pois}\left(2 * \Phi\left(\tilde{X}_j\right)\right), j = 5, 6, 7, 8$, generated from Poisson distribution;
- $X_j = \Phi\left(\tilde{X}_j\right), j = 9, 10$, where Φ is the cumulative distribution function of the standard Normal distribution.

1.2 Risk Models

The first model for the event time T followed a classical Cox Proportional Hazards Model:

$$\lambda(t \mid \mathbf{X}) = 0.6 \exp\left(b_0 + \beta^\top \mathbf{X}\right), b_0 = -3, \beta = 0.5*(1, -1, -2, -2, 1, -1, -2, -2, 1, -2)^\top$$

where $\lambda(t \mid \mathbf{X})$ is the hazard for the disease of interest. Additionally, to consider a setting with a more complex relationship between event status δ and predictors $\mathbf{X}$, we defined a time-varying relative risk model with interaction effects:

$$\lambda(t \mid \mathbf{X}) = 0.05 \cdot \exp\left(\frac{b_0 + \beta^\top \mathbf{X} + \mathbf{X}^\top \mathbb{B} \mathbf{X}}{t+1}\right), b_0 = -30, \beta = 0.5*(1, -1, -2, -2, 1, -1, -2, -2, 1, -2)^\top,$$

where $\mathbb{B}$ is a matrix with elements $b_{i,j} = \left\{1 - (-1)^{i+j} \times 3 + \mathrm{I}(i = j) \times 1.3\right\}/2$.

1.3 Outcomes and Surrogates

The censoring time C was generated from the discrete uniform distribution, as follows:

$$C_i = \left\lfloor \tilde{C}_i \right\rfloor, \tilde{C}_i \sim \mathrm{Unif}(20, 24).$$

The final current status δ was constructed, by definition, as $\delta = \mathrm{I}\{T \leq C\}$, with β and $\mathbb{B}$ chosen to yield a censoring rate about 50%.

From the true event time T, we generated the surrogate data S involving three longitudinal count variables $\mathbf{W}(t)$) to be used as inputs to derive the surrogates S through MAP. The last component $H(t)$ emulated the EHR utilization variable:

$$H(t) \sim \mathrm{Pois}\left(\lambda_{1,t}\right), \lambda_{1,t} = a_{h,t} I(t < T) + b_{h,t} I(t \geq T) - c_{h,t} I(t \geq T + 2))$$

$$W_1(t) \sim \mathrm{Pois}\left(\lambda_{2,t} H(t)\right), \lambda_{2,t} = a_{1,t} I(t < T) + b_{1,t} I(t \geq T) - c_{1,t} I(t \geq T + 2));$$

$$W_2(t) \sim \mathrm{Pois}\left(\lambda_{3,t} H(t)\right), \lambda_{3,t} = a_{2,t} I(t < T) + b_{2,t} I(t \geq T) - c_{2,t} I(t \geq T + 2)).$$

$$W_3(t) \sim \mathrm{Pois}\left(\lambda_{4,t} H(t)\right), \lambda_{4,t} = a_{3,t} I(t < T) + b_{3,t} I(t \geq T) - c_{3,t} I\left(t \geq T + 2\right)$$

$$W_4(t) \sim \mathrm{Pois}\left(\lambda_{5,t} H(t)\right), \lambda_{5,t} = a_{4,t} I(t < T) + b_{4,t} I(t \geq T) - c_{4,t} I\left(t \geq T + 2\right)$$

$$W_5(t) \sim \mathrm{Pois}\left(\lambda_{6,t} H(t)\right), \lambda_{6,t} = a_{5,t} I(t < T) + b_{5,t} I(t \geq T) - c_{5,t} I\left(t \geq T + 2\right)$$

Vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}$ has some correlation. We first generate a correlation matrix $\mathbf{M}$ and then generate $\mathbf{a}, \mathbf{b}, \mathbf{c}$

Then compute the co-occurrence matrix of and obtain the corresponding embeddings matrix through the PCA method[1].

1.4 S and Y

We derived S from MAP[2] using $\mathbf{W}(C)$ at censoring time under the bi-variate Poisson mixture model or

$$\mathcal{S}(t) \sim I(t < T)\,\mathrm{Beta}(1,3) + I(t \geq T)\,\mathrm{Beta}(1 + (t-T)/2, 1)$$