Simulation data generation

Ayden Zhao

No Institute Given

1. Generate $\tilde{\mathrm{X}}$ representation for each patient

We design two synthetic EHR datasets with varying levels of complexity in the predictor-outcome relationship. In all two datasets, we generate $N = 10000$ samples, with $n = 2000$ of these being annotated. For every patient, we first generate a multivariate Gaussian p-vector $\mathbf{X}(p = 10)$, with exchangeable correlation as follows:

$$\mathbf{X} \sim N(\mathbf{0}_p, \sigma^2_{P\times P})$$

Every element of $\mathbf{X}$ is corresponding to one EHR concept that we're interested in, which represent the patient's baseline probability of occuring corresponding disease.
σ^2 is the covariance matrix of all p EHR concepts. To calculate σ^2, we first obtain each EHR concept's semantic embedding e from co-occurrence matrix using SPPMI-SVD algorithm. Then the covariance matrix can be computed by:

$$\sigma^2(i,j) = e_i * e_j$$

where e_i and e_j are the codified semantic embedding of i_{th} and j_{th} EHR concept. Here for the simulation data, we simply just assume the codified concepts' semantic embedding $\{e_i, i \in [1, 10]\}$.

2. Derive the event time T of each patient from X

Secondly, we derive the event time T from $\mathbf{X}$. Each dataset is characterized by one of two different risk (hazard) models for the event time T. The first model we use to define the event time T is a classical Cox Proportional Hazards Model:

$$\lambda(t|\mathbf{X}) = 0.05 \cdot 12 \cdot \exp(b_0 + \boldsymbol{\beta}^\top \mathbf{X}), b_0 = -3, \beta = 0.5 * (1, -1, -2, -2, 1, -1, -2, -2, 1, -2)^\top$$

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

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

$$B = A + \frac{1}{2}\mathrm{diag}(1, 3, 1, 3, 1, 3, 1, 3, 1, 3) + J_{10}, A = \{-\frac{1 + 3(-1)^{i+j}}{2}\}_{ij}, J_{10} = \mathbb{1}\mathbb{1}'$$

3. Generate longitudinal count variable S for each patient

The most important step is to generate longitudinal count variables $\mathbf{S}$ for each patient. $\mathbf{S}$ for each patient is a $P \times T$ matrix, where P is the number of EHR concepts and T is the event time of patient. $S(i,t) = S_i(t)$ is the i_{th} EHR concept's mentioned times in the medical notes at t_{th} time spot.
From the defined T and $\mathbf{X}$, we generate $\mathbf{S}$ as follows:
We first generate $S_{base}(t)$ emulates the EHR utilization variable:

$$S_{base}(t) \sim Pois(\lambda_{0,t}),\ \lambda_{0,t} = 0.5I(t < T) + 4I(T \le t < T + 1) + 2I(t \ge T + 1));$$

$S_i(t)$ represents the i_{th} concepts' numbers of mention in medical notes:

$$S_i(t) \sim Pois(\lambda_{i,t} S_{base}(t)),\ \lambda_{i,t} = \Phi(\mathbf{X}_i)(0.05I(t < T) + 3I(T \le t < T + 1) + 2I(t \ge T + 1)));$$

And the final longitudinal count variables S for each patient is:

$$\mathbf{S} = (S_1(t), S_2(t), S_3(t), S_4(t), S_5(t), S_6(t), S_7(t), S_8(t), S_9(t), S_{10}(t))$$

4. Generate censoring time C and label Y

The censoring time C is then generated from the uniform distribution, as follows:

$$C_i = \lfloor \tilde{C}_i \rfloor, \ \tilde{C}_i \sim \text{Unif}(8, 34).$$

After T and C are determined, the final censoring status δ is defined, by definition, as $\delta_i = \text{I}\{T_i \leq C_i\}$, with $\boldsymbol{\beta}$ and B chosen to yield a censoring rate of approximately 50%. $\mathbf{S}$ is intercepted from $\mathbf{S}_{P\times T}$ to $\mathbf{S}_{P\times C}$.

5. Example

```
"","ID","Y","T","S.1","S.2","S.3","S.4","S.5","S.6","S.7","S.8","S.9","S.10"
"100",6,0,0,0,0,0,0,0,0,0,0,0,0
"101",6,0,1,0,0,0,0,0,0,0,0,0,0
"102",6,0,2,0,0,0,0,0,0,0,0,0,0
"103",6,0,3,0,0,0,0,0,0,0,0,0,0
"104",6,0,4,0,0,0,0,0,0,0,0,0,0
"105",6,0,5,0,0,0,0,0,0,0,0,0,0
"106",6,0,6,0,0,0,0,0,0,0,0,0,0
"107",6,0,7,0,0,0,0,0,0,0,0,0,0
"108",6,0,8,0,0,0,0,0,0,0,0,0,0
"109",6,1,9,13,5,6,7,7,10,11,18,4,4
"110",6,1,10,8,4,3,4,7,12,14,11,2,4
"111",6,1,11,3,0,4,2,3,6,6,3,4,2
"112",6,1,12,0,3,2,0,3,2,10,7,2,4
"113",6,1,13,2,1,0,1,1,1,2,2,0,0
"114",6,1,14,1,1,1,2,0,0,4,2,0,1
"115",6,1,15,4,1,2,4,3,5,4,8,4,2
"116",6,1,16,3,0,4,1,2,9,9,9,4,1
"117",6,1,17,7,0,5,2,1,2,4,7,1,0
"118",6,1,18,0,0,0,0,0,0,0,0,0,0
```

Fig. 1. Example for Generated Simulation Data