Cox model with distributed data using the WebDISCO algorithm

Estimates and confidence intervals for the parameters of the Cox model $\lambda(t|Z) = \lambda_0(t)\exp(\beta^T Z)$ are obtained by making use of the Newton-Raphson algorithm using local gradients and Hessian matrices. This allows for the recreation of the estimates and confidence intervals from the centralised setting. For the procedure, all nodes must have the same predictors.

The initial step consists of each node sharing all their unique event times to the coordination center. The global server then identifies and consolidates all unique event times across all sites. Each site then initializes parameters that remain constant throughout the process. Next, the coordination center initializes the first beta with an estimate of their choice.

The iterative process starts with each site computing data aggregates. These aggregates are then used by the central server to calculate gradients and Hessian matrices, which in turn are used to update the parameter estimates. This iterative process continues until convergence is achieved (which yields the centralised estimate).

In the following procedure, k is the site number, and M represents the number of sites ($k = 1, 2, \ldots, M$). N^k is the number of observations in a site, p is the number of predictors, i represents the event time number, and D represents the total number of distinct event time ($i = 1, 2, \ldots, D$). r and q are the indices of the element in the parameter vector β (r and $q = 1, 2, \ldots, p$), z^l is the variable z (predictors) for an individual subject l.

Example.

Suppose the following dataset at node k, with 5 observations ($N = 5$) and 2 predictors ($p = 2$):

Id	time	status	age	sex
1	3	1	42	1
2	6	0	38	1
3	11	1	37	2
4	11	1	51	1
5	14	1	36	2

where time represents the time of event occurrence or censoring, status indicates whether the event has occurred (1) or the data is censored (0) and age and sex are the predictors.

Data node (initial phase)

1. Each node identifies unique event times and saves this data to a CSV file. Each site also computes their local Cox models to obtain local betas and variance-covariance matrices, saving them to a CSV file along with the number of observations N^k. All files are then sent to the coordinating node.

Example (continued).

Estimates for the Cox proportional hazard model can be found using functions such as *coxph* in R, and we obtain:

$$\hat{\beta}^k = \begin{bmatrix} -0.0782 \\ -2.24456 \end{bmatrix}$$

The following data is **shared** to the coordinating node:

time
3
11
14

$$\hat{\beta}^k = \begin{bmatrix} -0.0782 \\ -2.24456 \end{bmatrix}, V_k = \begin{bmatrix} 0.0378 & 0.496 \\ 0.496 & 8.22 \end{bmatrix}, \text{and } N^k = 5.$$

Coordinating node (initial phase)

2. The coordinating node gathers all unique times from every site, generates a list of unique times across all sites, then sends it back to the data nodes. The global server also computes the inverse variance weighted initial estimator, $\hat{\beta}_0 = \left(\sum_{k=1}^{M} V_k^{-1} \hat{\beta}^k\right) / \sum_{k=1}^{M} V_k^{-1}$, and saves it to a CSV file.[1]

Data node (first iteration)

3. Each node initialises the following parameters: R_i^k, a list containing the IDs of the at-risk subjects for all times i, D_i^k, a list containing the IDs of subjects that had an event at time i, and $\sum_{l \in D_i^k} z_r^l$, the sum for a component r of the predictor for subjects with an event at time i. Each node also calculates $|D_i^k|$, the number of events for each time i. Both $|D_i^k|$ and $\sum_{l \in D_i^k} z_r^l$ are saved to a CSV file.

Example (continued).

The following data is computed but **not shared** to the coordinating node:

i	R_i^k
1	1,2,3,4,5
2	3,4,5
3	5

i	D_i^k
1	1
2	3,4
3	5

[1] Duan, R., Luo, C., Schuemie, M. J., Tong, J., Liang, C. J., Chang, H. H., ... & Chen, Y. (2020). Learning from local to global: An efficient distributed algorithm for modeling time-to-event data. *Journal of the American Medical Informatics Association*, *27*(7), 1028-1036.

The following data is **shared** to the coordinating node:

i	$\lvert D_i^k \rvert$
1	1
2	2
3	1

i	$\sum_{l \in D_i^k} z_r^l$	
	r = 1	***r = 2***
1	42	1
2	88	3
3	36	2

4. Then, each data node computes aggregated statistics used for the gradient and Hessian:

- $\sum_{l \in R_i^k} \exp(\boldsymbol{\beta^T z^l})$, a $D \; x \; 1$ matrix
- $\sum_{l \in R_i^k} \mathbf{z_q^l} \exp(\boldsymbol{\beta^T z^l})$, a $D \; x \; p$ matrix
- $\sum_{l \in R_i^k} \mathbf{z_r^l z_q^l} \exp(\boldsymbol{\beta^T z^l})$, a $p \; x \; p \; x \; D$ matrix

These quantities are sent to the coordinating center using CSV files.

Example (continued).

Suppose the coordinating node shared $\hat{\beta} = \begin{bmatrix} -0.05 \\ -2.5 \end{bmatrix}$.

For the first event time $(i = 1)$:

$$\sum_{l \in R_i^k} \exp(\beta^T z^l) = \exp\left([-0.05 \quad -2.5] * \begin{bmatrix} 42 \\ 1 \end{bmatrix}\right) + \exp\left([-0.05 \quad -2.5] * \begin{bmatrix} 38 \\ 1 \end{bmatrix}\right) + \cdots$$

$$\sum_{l \in R_i^k} z_q^l \exp(\beta^T z^l) = [42 \quad 1] \exp\left([-0.05 \quad -2.5] * \begin{bmatrix} 42 \\ 1 \end{bmatrix}\right) + [38 \quad 1] \exp\left([-0.05 \quad -2.5] * \begin{bmatrix} 38 \\ 1 \end{bmatrix}\right) + \cdots$$

$$\sum_{l \in R_i^k} z_r^l z_q^l \exp(\beta^T z^l) = \begin{bmatrix} 42^2 & 42 * 1 \\ 1 * 42 & 1^2 \end{bmatrix} \exp\left([-0.05 \quad -2.5] * \begin{bmatrix} 42 \\ 1 \end{bmatrix}\right)$$

$$+ \begin{bmatrix} 38^2 & 38 * 1 \\ 1 * 38 & 1^2 \end{bmatrix} \exp\left([-0.05 \quad -2.5] * \begin{bmatrix} 38 \\ 1 \end{bmatrix}\right) + \cdots$$

The following quantities are **shared** to the coordinating node:

i	$\sum_{l \in R_i^k} \exp(\beta^T z^l)$
1	0.031
2	0.0086
3	0.0011

i	$\sum_{l \in R_i^k} z_q^l \exp(\beta^T z^l)$	
	q = 1	***q = 2***
1	1.295	0.0331
2	0.406	0.0108
3	0.0401	0.0022

i		$\sum_{l\in R_i^k} z_r^l z_q^l \exp(\beta^T z^l)$	
		q = 1	***q = 2***
1	***r = 1***	55.02	1.374
	r = 2	1.374	0.037
2	***r = 1***	19.56	0.485
	r = 2	0.485	0.015
3	***r = 1***	1.443	0.080
	r = 2	0.080	0.0045

Coordinating node (first iteration only)

5. The global server aggregates $\sum_{k=1}^{M}\sum_{i=1}^{D}\sum_{l\in D_i^k} z_r^l$, which is the sum of all predictors for subjects that had an event across all sites, and calculates $\sum_{k=1}^{M}|D_i^k|$, the total number of events at each event time across all sites.

Coordinating node (all iterations)

6. The coordinating center computes the global gradient and Hessian:

$$l_r'(\beta) = \sum_{k=1}^{M}\sum_{i=1}^{D}\sum_{l\in D_i^k} z_r^l - \sum_{i=1}^{D}\left(\sum_{k=1}^{M}|D_i^k|\right)\frac{\sum_{k=1}^{M}\sum_{l\in R_i^k} z_r^l \exp(\beta^T z^l)}{\sum_{k=1}^{M}\sum_{l\in R_i^k}\exp(\beta^T z^l)}$$

$$l_{r,q}''(\beta) = -\sum_{i=1}^{D}\left(\sum_{k=1}^{M}|D_i^k|\right)\left\{\frac{\sum_{k=1}^{M}\sum_{l\in R_i^k} z_r^l z_q^l \exp(\beta^T z^l)}{\sum_{k=1}^{M}\sum_{l\in R_i^k}\exp(\beta^T z^l)} - \frac{\sum_{k=1}^{M}\sum_{l\in R_i^k} z_r^l \exp(\beta^T z^l)}{\sum_{k=1}^{M}\sum_{l\in R_i^k}\exp(\beta^T z^l)}\frac{\sum_{k=1}^{M}\sum_{l\in R_i^k} z_q^l \exp(\beta^T z^l)}{\sum_{k=1}^{M}\sum_{l\in R_i^k}\exp(\beta^T z^l)}\right\}$$

It then computes the Newton-Raphson iteration $\beta^{\tau} = \beta^{\tau-1} - [l''(\beta^{\tau-1})]^{-1} l'(\beta^{\tau-1})$ and sends the new estimate to the data nodes.

Steps **4.** and **6.** are repeated manually until convergence.

7. For all iterations except the first, the coordinating node computes the confidence intervals. The (estimated) covariance matrix of the coefficients is $\hat{\Sigma} = -\left(l''(\beta^{\tau-1})\right)^{-1}$.

Lower and upper bounds for the confidence intervals of the model parameters are also calculated at the coordinating node:

$$CI(\beta) = \left[\beta^{\tau-1} \pm z_{\frac{\alpha}{2}}\sqrt{diag(\hat{\Sigma})}\right]$$

The outputs of the procedure are the coefficients, the upper and lower bounds of the confidence intervals for the exponential of the model parameters, the standard error and the p values, for the previous iteration.