Cox model with distributed data

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 number of predictors. The total number of observations (individuals) is N and the number of predictors is p.

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.

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). !!!définir les symboles

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 all the unique times where an event has happened. This data is saved as a *csv* file and sent to the coordination node. Each site computes its local Cox models to obtain local betas, and saves it to a csv file, along with the number of observations.

Example (continued).

Beta can be found using functions such as coxph in R, and we obtain:

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

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

time
3
11
14

$\beta = \begin{bmatrix} -0.0782 \\ -2.24456 \end{bmatrix}$ et $N = 5$.

Coordinating node (initial phase)

2. The coordinating node gathers all unique times from every site, compares and generates a list of all unique times across all sites, then sends that list back to the data nodes. The global server also computes a first beta estimate, $\beta_0 = \left(\sum_{k=1}^{K} N^{(k)}\, \beta^{(k)}\right) / \sum_{k=1}^{K} N^{(k)}$, and writes it in a csv.

Data node (first iteration)

3. Each node initialises the parameters needed. These parameters are: R_i^k, a list containing all of the id of the subjects at risk for all times, D_i^k, a list containing the id of subjects where an even has happened, and $\sum_{l \in D_i^k} z_r^l$, the sum of all predictors for subjects that had an event. This data is saved as a *csv* file.

Example (continued).

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

$\mathbf{D_i^k}$
1
3,4
5

$\sum_{\mathbf{l \in D_i^k}} \mathbf{z_r^l}$	
42	1
88	3
36	2

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

$\mathbf{R_i^k}$
1,2,3,4,5
3,4,5
5

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

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

Where D is the number of event times (in this case, 3).

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

Example (continued).

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

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

$$\sum_{l \in R_i^k} \exp(\boldsymbol{\beta}^T \mathbf{z}^l)$$
$$\sum_{l \in R_i^k} \mathbf{z_q^l} \exp(\boldsymbol{\beta}^T \mathbf{z}^l)$$
$$\sum_{l \in R_i^k} \mathbf{z_r^l z_q^l} \exp(\boldsymbol{\beta}^T \mathbf{z}^l)$$

The exported *csv* will share the following table from node k:

Pour le premier temps:

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

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

Pour le premier temps:

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

t	$\sum_{l\in R_i^k} \mathbf{z_q^l} \exp(\boldsymbol{\beta^T z^l})$	
1	1.295	0.0331
2	0.406	0.0108
3	0.0401	0.0022

Pour le premier temps:

$$\sum_{l\in R_i^k} \mathbf{z_r^l z_q^l} \exp(\boldsymbol{\beta^T z^l}) = \begin{bmatrix} 42^2 & 42*1 \\ 1*42 & 1^2 \end{bmatrix} \exp\left([-0.05 \quad -2.5] * \begin{bmatrix} 42 \\ 1 \end{bmatrix}\right) + \begin{bmatrix} 38^2 & 38*1 \\ 1*38 & 1^2 \end{bmatrix} \exp\left([-0.05 \quad -2.5] * \begin{bmatrix} 38 \\ 1 \end{bmatrix}\right) + \cdots$$

t	$\sum_{l\in R_i^k} \mathbf{z_r^l z_q^l} \exp(\boldsymbol{\beta^T z^l})$	
1	55.02	1.374
	1.374	0.037
2	19.56	0.485
	0.485	0.015
3	1.443	0.080
	0.080	0.0045

Coordinating node (first iteration only)

5. The coordinating node calculates $|D_i^k|$, which is the number of events for every event time. The global server also 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 all subjects that had an event, all sites together.

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 **3.** and **4.** are repeated until convergence or a fixed number of times (T).

Coordinating node (last iteration T)

7. In addition to a normal iteration, the coordinating node computes the confidence intervals.

HOW?