ODACH: A One-shot Distributed Algorithm for Cox model with Heterogeneous Multi-center Data

Authors:

Chongliang Luo¹, Rui Duan², Yong Chen¹

¹Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA, USA

²Department of Biostatistics, Harvard T.H. School of Public Health, Boston, MA, USA

ABSTRACT

Objective:

We developed and evaluated a privacy-preserving One-shot Distributed Algorithm for Cox model to analyze multi-center time-to-event data without sharing patient-level information across sites, while accounting for heterogeneity across sites by allowing site-specific baseline hazard functions and feature distributions.

Materials and Methods:

We constructed a surrogate likelihood function to approximate the Cox log partial likelihood function which is stratified by site, using patient-level data from a single site and aggregated information from other sites. The ODAC estimator was obtained by maximizing the surrogate likelihood function. We evaluated and compare the performance of ODACH with meta-analysis by extensive numerical studies.

Results:

The simulation study showed that ODACH provided estimates close to the pooled estimator, which is obtained by directly analyzing patient-level data from all sites via a stratified Cox model. The relative bias was <1% across all scenarios. As a comparison, the meta-analysis estimator, which was obtained by the inverse variance weighted average of the site-specific estimates, had substantial bias when the event rate is <5%, with the relative bias reaching 12% when the event rate is 1%.

Conclusions:

ODACH is a privacy-preserving and communication-efficient method for analyzing multi-center time-to-event data, which allows the baseline hazard functions as well as the distribution of covariate variables to vary across sites. It provides estimates that is close to the pooled estimator and substantially outperforms the meta-analysis estimator when the event is rare. It is thus extremely suitable for studying rare events with heterogeneous baseline hazards across sites in a distributed manner.

Keywords: Stratified Cox proportional hazard model, Distributed algorithm, Electronic health record, Heterogeneity.

Word count: 2295

INTRODUCTION

Real-world data (RWD) such as electronic health records (EHRs) and medical claims are drawing more attention in generating real-world evidence to support healthcare decision making [1-3]. The past decade has seen an increasing number of clinical research networks, curating and using various EHRs and medical claims data. For example, the Observational Health Data Sciences and Informatics (OHDSI) network is an international network that cover more than half a billion patient records in a common data model (CDM) [5]. Another example is the national Patient-Centered Clinical Research Network (PCORnet), a network covering more than 100 million patients in the United States [6]. These large data consortia provide platforms and tools to integrate heterogenous RWD from a diverse range of healthcare organizations. Multicenter analyses using RWD from these clinical research networks have been increasing rapidly in large part because they can improve the generalizability of the study results due to increased sample size.

Despite the benefits of multicenter analyses, two major challenges exist for multi-site data integration. First, the direct sharing of patient-level data across institutions may be prohibited, as the individual patient-level data are protected by privacy regulations such as the Health Insurance Portability and Accountability Act (HIPAA) or General Data Protection Regulation (GDPR). Hence, many research networks, such as OHDSI, adopt a federated model in which the patient-level data are stored at local institutions and only aggregated information are shared [6-8]. Second, data from different sites are often heterogeneous due to different patient characteristics, data quality, and other unobserved site-level features. Assuming a common statistical model across all sites may result in biased estimation and poor prediction.

Survival analyses are frequently used in biomedical research to model the time to a specific event of interest. To conduct multi-center survival analyses without sharing patient-level data, a common and convenient approach is the meta-analysis, e.g. a weighted average of the local estimates from each site. However, when the outcomes or exposures are rare, or the sample sizes are small in some sites, the accuracy of the meta-analysis may be poor [19,29]. Other distributed algorithms were developed to obtain more accurate results. For example, WebDISCO (a web service for distributed Cox model learning) [14] were developed for conducting logistic regression and Cox regression respectively. Despite the fact that these algorithms can provide identical results to the analysis using the pooled individual-level data (lossless), they are communication intensive due to the iterative nature of the algorithms which requires multiple rounds of communications across site. To balance the communication efficiency and estimation accuracy, Shu et al. [30] proposed a lossless one-shot algorithm for stratified Cox model which can only include one binary covariate in the model. Wang et al. [31] proposed a divide and conquer approach, which aims to reduce the computational challenge when the sample size is extremely large. In addition, Duan et al. [29] proposed a One-shot Distributed Algorithm for Cox model (ODAC) based on the surrogate likelihood approach which relies on patient-level data from a single site and aggregated data from other sites. The algorithm only requires aggregated data twice from other sites but obtains close estimation as if using patient-level data from all sites.

Reviewing the mentioned approaches, most of them are based on the Cox proportional hazards model, but few of them take the data heterogeneity across sites into account. Specifically, in the multi-center survival analysis, the distribution of covariates is likely to be different across sites as

patients might come from different sub-populations. In addition, the baseline hazard functions might also be heterogenous across sites. Assuming a Cox model with a common baseline hazard function may lead to biased results. In this article, we propose a distributed algorithm which account for site-level heterogeneities on covariate distributions and baseline hazard functions, namely the One-shot Distributed Algorithm for Cox model with Heterogeneity (ODACH). Compared to the previous work of ODAC that assumes common baseline hazard function across sites, the assumption of heterogeneous baseline hazard functions in ODACH is more flexible and practical in real-world data. Moreover, the constructed surrogate likelihood does not require one extra round of communication regarding the risk set in each site as in ODAC. As a result, ODACH is more communication efficient over ODAC. We show that our proposed algorithm is one-shot and achieves high accuracy (small bias) in a simulation study.

METHODS

Stratified Cox Proportional Hazard Model

To account for site-specific baseline hazards, we introduce the stratified Cox model for time-to-event outcome. Suppose we have study subjects from K different clinical sites and denote n_j to be the number of subjects in the j -th site. Denote the total number of subjects to be $N = \sum_{j=1}^K n_j$. For the i -th subject in the j -th site, we observe $\{T_{ij}, \delta_{ij}, x_{ij}\}$, where T_{ij} is the observed time to event, x_{ij} is a p -dimensional covariate variable, and $\delta_i = 0$ indicates censoring and $\delta_i = 1$ indicates an event. Assume that the j -th site has its own baseline hazard function $\lambda_j(t)$, the Cox proportional hazard model assumes that the hazard of the i -th subject in the j -th site having the event at time t follows

$$\lambda(t|x_{ij}) = \lambda_j(t) \exp(\beta^T x_{ij})$$

where β contains the log hazard ratios. We assume that β is the same across all sites, i.e. common effects of the covariates on the time-to-event across sites. For the j -th site, the log Cox partial likelihood function is

$$L_j(\beta) = \frac{1}{n_j} \sum_{i=1}^{n_j} \delta_{ij} \log \frac{\exp(\beta^T x_{ij})}{\sum_{s \in R_j(T_{ij})} \exp(\beta^T x_{sj})}, \quad (1)$$

where $R_j(t)$ is the risk set in site j at time t defined as $R_j(t) = \{i; T_{ij} \geq t\}$, which contains all the subjects in site j that have not experienced an event or been censored at time t . Notice that the partial likelihood does not involve the baseline hazard $\lambda_j(t)$. To deal with the heterogeneity of baseline hazard functions, we can construct a stratified log Cox partial likelihood function

$$L(\beta) = \frac{1}{N} \sum_{j=1}^K n_j L_j(\beta). \quad (2)$$

The common effect β can be estimated by maximizing the stratified function (1), i.e. $\hat{\beta} = \operatorname{argmax}_{\beta} L(\beta)$. We call it the pooled estimator, as it requires all the data to be pooled together.

A One-Shot Distributed Algorithm for Cox Model with Heterogeneity

In practice, the patient-level data are often difficult to be transferred across sites. Hence the pooled estimate $\hat{\beta}$ could be hard to obtain. Inspired by the surrogate likelihood approach developed in [9,13,20], we aim to construct a proxy of the stratified Cox partial likelihood function (1), which we call a surrogate likelihood function, using only summary-level information from other sites. Define $\bar{\beta}$ to be an initial value that is close to the true value of the parameter β , and we only have access to the patient-level data of a leading site (e.g. the first site), we propose to construct the following surrogate likelihood function

$$\tilde{L}(\beta) = L_1(\beta) + \langle \nabla L(\bar{\beta}) - \nabla L_1(\bar{\beta}), \beta \rangle + \frac{1}{2} (\beta - \bar{\beta})^T \{ \nabla^2 L(\bar{\beta}) - \nabla^2 L_1(\bar{\beta}) \} (\beta - \bar{\beta}), \quad (3)$$

where $L_1(\beta)$ is the local likelihood function defined in (1), ∇ and ∇^2 denote the first and second order gradients of a function (explicit forms of $\nabla L_j(\bar{\beta})$, $\nabla L(\bar{\beta})$, $\nabla^2 L_j(\bar{\beta})$ and $\nabla^2 L(\bar{\beta})$ can be found in the Appendix). Intuitively, the above surrogate likelihood function (3) modifies the local likelihood function $L_1(\beta)$ to approximate the stratified likelihood (2), with the modification being the first and second-order terms, i.e. $\langle \nabla L(\bar{\beta}) - \nabla L_j(\bar{\beta}), \beta \rangle$ and $\frac{1}{2}(\beta - \bar{\beta})^T \{\nabla^2 L(\bar{\beta}) - \nabla^2 L_j(\bar{\beta})\}(\beta - \bar{\beta})$. By sharing the second-order gradients, our method allows each site to have different covariate distributions. The surrogate estimator is obtained by maximizing the surrogate likelihood (3), i.e. $\tilde{\beta} = \operatorname{argmax}_{\beta} \tilde{L}(\beta)$.

In the surrogate likelihood function (3), $\nabla L(\bar{\beta})$ and $\nabla^2 L(\bar{\beta})$ can be calculated distributively by $\nabla^r L(\bar{\beta}) = \frac{1}{N} \sum_{j=1}^K n_j \nabla^r L_j(\bar{\beta})$, for $r = 1, 2$. Since $\nabla L_1(\bar{\beta})$ and $\nabla^2 L_1(\bar{\beta})$ are available from the leading site, it only requires other collaborative sites to calculate and transfer $\nabla L_j(\bar{\beta})$ and $\nabla^2 L_j(\bar{\beta})$, $j = 2, \dots, K$. These gradients are all aggregated information, hence patient-level information is protected. Regarding the initial value $\bar{\beta}$, we recommend to use the inverse variance weighted average of the estimates obtained by fitting a Cox model at each site, i.e.,

$$\bar{\beta} = \left(\sum_{j=1}^K \hat{V}_j^{-1} \right)^{-1} \sum_{j=1}^K \hat{V}_j^{-1} \hat{\beta}_j \quad (4)$$

where $\hat{\beta}_j = \operatorname{argmax}_{\beta} L_j(\beta)$ is estimator of Cox model fitted on data in the j -th site, and $\hat{V}_j$ is the estimated variance of $\hat{\beta}_j$. We summarize the ODACH algorithm below and also provide a schematic illustration in Figure 1. Notice in the above derivation, we assume the first site is the leading site when constructing the surrogate likelihood. In practice if all the sites can serve as the leading site, we recommend using the largest site as the leading site. Alternatively, after the

derivatives $\nabla^r L_j(\bar{\beta}), r = 1, 2, j = \dots, K$ have been shared across sites, each site can serve as the leading site and obtain its own surrogate estimate. These surrogate estimates can be further synthesized to obtain more accurate estimation.

Algorithm ODACH

(1) Initialization

In Site $k = 1$ to K ,

do

Fir a Cox regression model and obtain the local estimate $\hat{\beta}_k$ and the variance estimate $\hat{V}_k$

broadcast $\hat{\beta}_k, \hat{V}_k$.

end

(2) Aggregated data communication

In Site $k = 1$ to K ,

do

obtain $\bar{\beta}$ using (5)

calculate and broadcast the gradients $\nabla L_j(\bar{\beta})$, and $\nabla^2 L_j(\bar{\beta})$

end

(3) Local surrogate estimator

In the leading site $k = 1$

do

construct surrogate likelihood $\tilde{L}(\beta)$ in (3)

obtain $\tilde{\beta}$ by maximizing $\tilde{L}(\beta)$.

Return $\tilde{\beta}$.

Figure 1: Schematic illustration of the ODACH algorithm. The first step is initialization where each site reports the local estimation of log hazard ratio ($\hat{\beta}_j$), and the corresponding variance estimate ($\hat{V}_j$). The leading site then compute the initial estimate $\bar{\beta}$ as the weighted average of all local estimates, and send it back to each site. In the second step, each site calculates and shares the local gradients $\nabla L_j(\bar{\beta})$ and $\nabla^2 L_j(\bar{\beta})$. In the third step, the leading site constructs a surrogate likelihood function $\tilde{L}(\beta)$ with these gradients, and obtains the surrogate estimate $\tilde{\beta}$.

SIMULATION STUDY

We evaluate the performance of the proposed ODAC estimator using simulated multi-site time-to-event data. A pooled dataset of $N=5,000$ subjects are generated based on Weibull proportional hazard model, where the baseline hazard follows a Weibull-distribution with varying scale and shape parameters. Specifically, the scale parameters range from 100 to 280 with equal space. The shape parameters range from 20 to 0.5 with equal space in the logarithm scale, see Figure 2 for an illustration. We generate 2 covariates from uniform distributions and the true log hazard ratios are set to be $\beta = (-1, 1)$. We set the event rate (number of cases over number of subjects) as 20%, 2% and 1% by generating censoring times following appropriate distributions. The pooled data are evenly distributed to $K=10$ clinical sites, with 500 subjects in each site. We apply three approaches to estimate the hazard ratios of the two covariates on the time-to-event outcome, i.e. pooled

stratifies Cox regression, the meta-analysis, and the proposed ODACH method. Since the pooled Cox regression estimator can be considered as a gold standard, the relative bias of meta-analysis and ODACH estimates to the pooled estimate are compared to demonstrate the advantage of ODACH. The simulation was replicated 200 times. For simplicity of illustration, we only present the results for estimation of coefficient β_2 and the results for the other coefficient are similar.

Figure 2. The baseline survival functions of the 10 sites in the simulated data. The varying hazard functions are Weibull functions with scale and shape parameters as listed.

The simulation results show that ODACH achieves better estimation performance than the meta-analysis estimator, especially when the event is rare. In Figure 3, ODACH obtains relative biases close to 0, meaning that it provides almost identical results to the pooled estimator, i.e. by stratified Cox model on the pooled dataset across all sites. When the event becomes more rare, meta-analysis estimator is observed to have larger bias. For example, when the event rate is 1%, the average relative bias is around -12% for the meta-analysis estimator, but only around -1% for ODACH

estimator. Moreover, the variation of the meta-analysis estimator is much larger compared with that of the ODACH estimator.

Figure 3. Boxplot of relative bias to the gold standard (stratified Cox model on the pooled dataset across all sites). The 2 methods compared in the plot are meta-analysis (meta) and One-shot Distributed Algorithm for Cox model with Heterogeneous baseline hazards (ODACH). The event rate varies from 20% to 1%, and under each setting, the boxplots are based on 200 replications of the simulation. The true effect size is 1.

DISCUSSION

We developed a privacy-preserving One-shot Distributed Algorithm for the Cox model to analyze Heterogeneous multi-center time-to-event data (ODACH). The proposed surrogate likelihood approach approximates the log partial likelihood of the stratified Cox model that uses the patient-level data from all the sites. The simulation study shows the surrogate estimation obtains closer estimation towards the pooled analysis compared to the meta-analysis, especially when the event

is rare. Compare to the existing One-shot Distributed Algorithm for Cox model (ODAC), ODACH allows the baseline hazard functions and covariate distributions to be site-specific, and hence is more flexible in application.

Real-world data are playing an increasing role in generating evidence to support healthcare decision making. Observational data such as EHR and medical claims data contains longitudinal information, which enables time-to-event modeling. The Cox model is one of the most commonly used models for time-to-event analysis, and has been widely applied in observational studies for treatment evaluation and risk factor identification. For example, Suchard et al. [32] conducted a systematic study to compare first-line antihypertensive drugs using a global network of administrative claims and EHR databases. The Cox model was used to study time-to-event outcomes regarding effectiveness and safety. Without sharing the patient-level data across databases, the individual estimates from each database are integrated by a meta-analysis approach. In such multi-center studies, our proposed distributed algorithm could be an alternative to the commonly used meta-analysis approach and have potential benefits in the scenario of rare events. The algorithm is implemented in the R package “pda” and is available on CRAN [33].

There are several directions of future work. Time-varying covariates or time-varying effects are sometimes encountered in time-to-event analyses [34,35]. The Cox model with these variants relaxes the proportional hazards assumption of the usual Cox model, but requires additional data for accurate estimation [36,37]. It is thus desired to develop distributed algorithm for the Cox model with time-varying covariates or time-varying effects in multi-center studies. Moreover, other survival models, such as the accelerated failure time (AFT) model [38] can be more suitable

than the Cox model in certain settings. A distributed algorithm for AFT model is currently under investigation and will be reported in the future. In addition, other sources of heterogeneity other than baseline hazard functions or distributions might exist, such as missing data patterns and site-specific effect sizes. Robust methods for handling different types of heterogeneity are needed to avoid potentially misleading results.

Acknowledgement

This work was supported partially through a Patient-Centered Outcomes Research Institute (PCORI) Project Program Award (ME-2019C3-18315). All statements in this report, including its findings and conclusions, are solely those of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute (PCORI), its Board of Governors or Methodology Committee.

REFERENCES

1. Shea, S. and G. Hripcsak, *Accelerating the use of electronic health records in physician practices*. New England Journal of Medicine, 2010. **362**(3): p. 192-195.
2. Sherman, R.E., et al., *Real-world evidence—what is it and what can it tell us*. N Engl J Med, 2016. **375**(23): p. 2293-2297.
3. Friedman, C.P., A.K. Wong, and D. Blumenthal, *Achieving a nationwide learning health system*. Science translational medicine, 2010. **2**(57): p. 57cm29-57cm29.
4. Cox, D.R., *Regression models and life-tables*. Journal of the Royal Statistical Society: Series B (Methodological), 1972. **34**(2): p. 187-202.
5. Hripcsak, G., et al., *Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers*. Studies in health technology and informatics, 2015. **216**: p. 574.
6. Fleurence, R.L., et al., *Launching PCORnet, a national patient-centered clinical research network*. Journal of the American Medical Informatics Association, 2014. **21**(4): p. 578-582.

7. Schuemie, M.J., et al., *Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data*. Proceedings of the National Academy of Sciences, 2018. **115**(11): p. 2571-2577.
8. Duke, J.D., et al., *Risk of angioedema associated with levetiracetam compared with phenytoin: Findings of the observational health data sciences and informatics research network*. Epilepsia, 2017. **58**(8): p. e101-e106.
9. Vashisht, R., et al., *Association of hemoglobin A1c levels with use of sulfonylureas, dipeptidyl peptidase 4 inhibitors, and thiazolidinediones in patients with type 2 diabetes treated with metformin: analysis from the observational health data sciences and informatics initiative*. JAMA network open, 2018. **1**(4): p. e181755-e181755.
10. Boland, M.R., et al., *Uncovering exposures responsible for birth season–disease effects: a global study*. Journal of the American Medical Informatics Association, 2017. **25**(3): p. 275-288.
11. Hripcsak, G., et al., *Characterizing treatment pathways at scale using the OHDSI network*. Proceedings of the National Academy of Sciences, 2016. **113**(27): p. 7329-7336.
12. Chen, Y., et al., *Regression cubes with lossless compression and aggregation*. IEEE Transactions on Knowledge and Data Engineering, 2006. **18**(12): p. 1585-1599.
13. Wu, Y., et al., *Grid Binary LOGistic REGression (GLORE): building shared models without sharing data*. Journal of the American Medical Informatics Association, 2012. **19**(5): p. 758-64.
14. Lu, C.-L., et al., *WebDISCO: a web service for distributed cox model learning without patient-level data sharing*. Journal of the American Medical Informatics Association. 2015. **22**(6): p. 1212-1219.
15. Wang, J., et al. *Efficient distributed learning with sparsity*. in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. 2017. JMLR. org.
16. Jordan, M.I., J.D. Lee, and Y. Yang, *Communication-efficient distributed statistical inference*. Journal of the American Statistical Association, 2019. **114**(526): p. 668-681.
17. Duan, R., et al. *ODAL: A one-shot distributed algorithm to perform logistic regressions on electronic health records data from multiple clinical sites*. in *PSB*. 2019. World Scientific.
18. Ohno-Machado, L., et al., *pSCANNER: patient-centered Scalable National Network for Effectiveness Research*. Journal of the American Medical Informatics Association, 2014. **21**(4): p. 621-626.
19. Duan, R., et al. 2020. Learning from electronic health records across multiple sites: A communication-efficient and privacy-preserving distributed algorithm. *Journal of the American Medical Informatics Association*, 27(3), pp.376-385.
20. Overhage, J.M., et al., *Validation of a common data model for active safety surveillance research*. Journal of the American Medical Informatics Association, 2011. **19**(1): p. 54-60.
21. Anand, S.S., et al., *Risk factors for myocardial infarction in women and men: insights from the INTERHEART study*. European heart journal, 2008. **29**(7): p. 932-940.
22. Lanus, F., et al., *Risk factors for acute myocardial infarction in Latin America: the INTERHEART Latin American study*. Circulation, 2007. **115**(9): p. 1067-1074.
23. Kokotailo, R.A. and M.D. Hill, *Coding of stroke and stroke risk factors using international classification of diseases, revisions 9 and 10*. Stroke, 2005. **36**(8): p. 1776-1781.
24. Forrest, C.B., et al., *PEDSnet: a national pediatric learning health system*. 2014. **21**(4): p. 602-606.

25. Shenkman, E., et al., *OneFlorida Clinical Research Consortium: linking a clinical and translational science institute with a community-based distributive medical education model*. Academic Medicine, 2018. **93**(3): p. 451.
26. Informatics, O.H.D.S.a. *The Book of OHDSI*. Chapter 20 OHDSI Network Research 2019 October 13, 2019]; Available from: <https://ohdsi.github.io/TheBookOfOhdsi/NetworkResearch.html>.
27. Network, P.D.R. 2013 October 13, 2019]; Available from: <https://querytool.pcornet.org/>.
28. PopMedNet. 2012 October 13, 2019]; Available from: <https://www.popmednet.org/>.
29. Duan, R., et al. Learning from local to global: An efficient distributed algorithm for modeling time-to-event data, *Journal of the American Medical Informatics Association*, Volume 27, Issue 7, July 2020, Pages 1028–1036, <https://doi.org/10.1093/jamia/ocaa044>.
30. Shu, D., Yoshida, K., Fireman, B.H. and Toh, S., 2020. Inverse probability weighted Cox model in multi-site studies without sharing individual-level data. *Statistical Methods in Medical Research*, 29(6), pp.1668-1681.
31. Wang, Y., Hong, C., Palmer, N., Di, Q., Schwartz, J., Kohane, I. and Cai, T., 2019. A fast divide-and-conquer sparse Cox regression. *Biostatistics*. 22: 2, pp. 381–401.
32. Suchard, M.A., Schuemie, M.J., Krumholz, H.M., You, S.C., Chen, R., Pratt, N., Reich, C.G., Duke, J., Madigan, D., Hripcsak, G. and Ryan, P.B., 2019. Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. *The Lancet*, 394(10211), pp.1816-1826.
33. Luo, C., et al., 2020. pda: Privacy-Preserving Distributed Algorithms. R package version 1.0.2. <https://CRAN.R-project.org/package=pda>
34. Therneau, T., Crowson, C. and Atkinson, E., 2013. Using time dependent covariates and time dependent coefficients in the cox model. *Red*, 2(1).
35. Zhang, Z., Reinikainen, J., Adeleke, K.A., Pieterse, M.E. and Groothuis-Oudshoorn, C.G., 2018. Time-varying covariates and coefficients in Cox regression models. *Annals of translational medicine*, 6(7).
36. Cai, Z. and Sun, Y. (2003). Local linear estimation for time-dependent coefficients in cox's regression models. *Scandinavian Journal of Statistics* 30, 93–111.
37. Tian, L., Zucker, D., and Wei, L. (2005). On the cox model with time-varying regression coefficients. *Journal of the American Statistical Association* 100, 172–183.
38. Wei, L.J., 1992. The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Statistics in Medicine*, 11(14-15), pp.1871-1879.

Appendix

1. The local first order gradient $\nabla L_j(\bar{\beta})$

$$\nabla L_j(\bar{\beta}) = \frac{1}{n_j} \sum_{i=1}^{n_j} \delta_{ij} \left\{ x_{ij} - \frac{\sum_{s \in R_j(T_{ij})} \exp(\bar{\beta}^T x_{sj}) x_{sj}}{\sum_{s \in R_j(T_{ij})} \exp(\bar{\beta}^T x_{sj})} \right\} \quad (\text{A.1})$$

2. The local second order gradient $\nabla^2 L_j(\bar{\beta})$

$$\nabla^2 L_j(\bar{\beta}) = \frac{1}{n_j} \sum_{i=1}^{n_j} \delta_{ij} \left[\frac{\left\{ \sum_{s \in R_j(T_{ij})} \exp(\bar{\beta}^T x_{sj}) x_{sj} \right\}^{\otimes 2}}{\left\{ \sum_{s \in R_j(T_{ij})} \exp(\bar{\beta}^T x_{sj}) \right\}^2} - \frac{- \sum_{s \in R_j(T_{ij})} \exp(\bar{\beta}^T x_{sj}) x_{sj}^{\otimes 2}}{\sum_{s \in R_j(T_{ij})} \exp(\bar{\beta}^T x_{sj})} \right] \quad (\text{A.2})$$

where $a^{\otimes 2}$ for a vector a denotes the outer product aa^T .

3. The global first order gradient $\nabla L(\bar{\beta})$

$$\nabla L(\bar{\beta}) = \frac{1}{N} \sum_{j=1}^K n_j \nabla L_j(\bar{\beta}) \quad (\text{A.3})$$

4. The global second order gradient $\nabla^2 L(\bar{\beta})$

$$\nabla^2 L(\bar{\beta}) = \frac{1}{N} \sum_{j=1}^K \nabla^2 L_j(\bar{\beta}) \quad (\text{A.4})$$