Les Cahiers du GERAD ISSN: 0711–2440

Semi$^+$-supervised learning under sample selection bias

D. Robatian,
M. Asgharian

G–2020–23–EIW09

April 2020

La publication de ces rapports de recherche est rendue possible grâce au soutien de HEC Montréal, Polytechnique Montréal, Université McGill, Université du Québec à Montréal, ainsi que du Fonds de recherche du Québec – Nature et technologies.

Dépôt légal – Bibliothèque et Archives nationales du Québec, 2020
– Bibliothèque et Archives Canada, 2020

The publication of these research reports is made possible thanks to the support of HEC Montréal, Polytechnique Montréal, McGill University, Université du Québec à Montréal, as well as the Fonds de recherche du Québec – Nature et technologies.

Legal deposit – Bibliothèque et Archives nationales du Québec, 2020
– Library and Archives Canada, 2020

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

Semi$^+$-supervised learning under sample selection bias

Damoon Robatian [a]
Masoud Asgharian [b]

[a] *GERAD & Department of Mathematics and Industrial Engineering, Polytechnique Montréal, Montréal (Québec) Canada, H3C 3A7*

[b] *Department of Mathematics and Statistics, McGill University, Montréal (Québec) Canada, H3A 2K6*

damoon.robatian@polymtl.ca
masoud.asgharian2@mcgill.ca

April 2020
Les Cahiers du GERAD
G–2020–23–EIW09

Abstract: *In time-to-event data analysis, the main object of interest is the time elapsed between the occurrence of two ordered events, say E_1, E_2. Sampling from the incident population, i.e., subjects who have experienced the incidence of E_1 before being sampled regardless of the occurrence of E_2, is the gold standard in follow-up studies. Yet often in practice, it is more feasible to sample from the prevalent population, i.e., subjects who have already experienced E_1, but not E_2. It is well known that the prevalent sampling design induces sample selection bias. Moreover, time-to-event data are usually subject to censoring which causes partial loss of information on a fraction of the subjects. Here, we discuss the inefficiency of the conventional learning methods due to ignoring sample selection bias and show how this problem can be avoided by properly incorporating the selection bias into the analysis. Arguments are backed by simulation studies.*

1 Introduction

Time-to-event is the output of interest in numerous disciplines spanning epidemiology, economics, econometrics, gerontology, and etc. It is defined as the amount of time elapsed from the occurrence of an initiating event until that of a second event called terminating event. Both events are pre-defined. For example, the initiating event might be birth, onset of a disease, or an aircraft's release, while the terminating event could be retirement, death, or the aircraft's phase-out, respectively. Time-to-event modelling is a ubiquitous problem, an evidence of which is the existence of multiple domains, such as survival analysis, reliability theory, event history analysis, duration modelling, etc., all with similar objectives. As a result, a vast variety of methods have been developed for this purpose. Survival analysis alone hosts a great deal of theory, a big portion of which is related to modelling potential associations between the time-to-event or an individual's survival time and a set of observed measurements for that individual. Naturally, any data-driven inference depends on characteristics of the training data. That is, any quality of the data, potentially affecting the outcome of the analysis, should be properly incorporated in the learning process; otherwise the algorithm's learnability, i.e., the ability to extract relevant information might be influenced negatively. Regarding time-to-event data, there are several points worth considering. One is that data may suffer from multiple types of *incompleteness*, ignoring which may cause serious issues. The gold standard in time-to-event data is to conduct follow-up studies on randomly selected cases from the *incident population*, i.e., subjects who have not experienced the initiating event before the study starts. Logistic or other constraints may, however, preclude the possibility of conducting incident cohort studies. A feasible alternative in such cases is to conduct a *cross-sectional prevalent cohort study* for which one recruits prevalent cases, that is, subjects who have already experienced the initiating event, but not the terminating event. When the interest lies in estimating the lifespan between the initiating and the terminating event, subjects may be followed prospectively either until the terminating event happens or they are lost to follow-up, whichever occurs first. This study design gives rise to two types of incompleteness: First, the response variable, being lifetime, is observed for some subjects while for others we only know that it is greater than some observed period, called censoring time. This type of incompleteness due to censoring is called *right censoring*. Learning from such data for prediction and generalization falls, roughly, into the realm of semi-supervised learning with the difference that there is partial information on subjects whose response is not observed, hence the name semi$^+$-supervised learning. Second, it is well known that prevalent cases have, on average, longer lifespans since longer survivors are more prone to be selected at the recruitment time. As such, a prevalent cohort comprises a biased, and non-random sample of the target population. An important feature of such data is that the response selection bias induces a bias on the feature space's sampling frame. This is so since certain feature values could be preferentially selected into the sample, being linked to the long-term survivors, who themselves are favored by the sampling mechanism. This systematically introduced bias comprises the second type of incompleteness, called sample selection bias. Here, we discuss different challenges in analysing and learning from such data. Particular attention is paid to the case where the chance of being selected into the sample is proportional to the survival time, the so-called length-biased sampling. Especially, we

consider the learning problem in (i) variable selection, and (ii) classification settings, and will conclude that the conventional approach for learning lacks efficiency and describe how this can be fixed.

2 Training data

Throughout this note, uppercase letters denote one-dimensional random variables, while bold uppercase denotes random vectors. For any subject i, we define the following: The response variable, i.e., time-to-event, represented by U_i. The length-biased variables will be marked with a tilde, e.g., $\tilde{U}_i$ refers to the length-biased survival time. Naturally, we assume that $U_i, \tilde{U}_i \geq 0$. Also, $\mathbf{Z}_i = (Z_{i_1}, Z_{i_2}, \ldots, Z_{i_d})$, with $d \geq 1$, is a vector of covariates. To distinguish biased covariates, we shall use the superscript "$*$". Realizations of random variables and vectors are shown via lowercase and bold lowercase, respectively. To avoid cumbersome notation, no distinction between unbiased and biased realizations is made. When applicable, regression coefficients are denoted by $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_d)^\top$. Nevertheless, $\boldsymbol{\theta}$ is used to indicate the vector of all parameters to be estimated, including the regression ones. Define T_i as the *current lifetime*, i.e., the time interval between the initiating event and sampling time. Similarly, the *residual lifetime*, denoted by R_i, is defined to be the time interval from the sampling until the terminating event. Therefore, $U_i = T_i + R_i$. Also, let C_i denote the *censoring time*, which is the time elapsed from the sampling of the subject until its possible censoring. One may observe only $R_i \wedge C_i = \min(R_i, C_i)$ due to possible censoring. Additionally, we define $X_i := T_i + (R_i \wedge C_i)$. A failure indicator δ_i is defined to be a Bernoulli random variable indicating whether a subject has failed or censored; that is, $\delta_i = \mathbb{1}_{\{R \leq C\}}$. Finally, the training data is assumed to be of the following form:

$$\tilde{S}_c = \{(\tilde{T}_i, \tilde{R}_i \wedge C_i, \delta_i, \mathbf{Z}_i^*) : i = 1, \ldots, n\},$$

with $\mathbf{Z}_i^* = (Z_{i_1}^*, \ldots, Z_{i_d}^*)$. Recall that $\tilde{S}_c$ is length biased (see Figure 1).

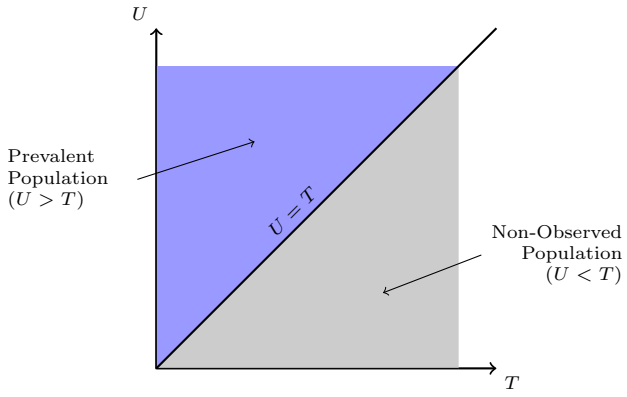

Figure 1: Prevalent vs. Incident Populations: The upper triangle depicts the prevalent population, while the incident population consists of both upper and lower triangles

3 Conditional vs. joint approaches

Consider the regression problem

$$U_i = f_{\boldsymbol{\beta}}(\mathbf{Z}_i) + \boldsymbol{\varepsilon}_i, \; i = 1, \ldots, n \tag{1}$$

where U_i is the response, $\mathbf{Z}_i \in \mathbb{R}^d$ a vector of covariates, $\boldsymbol{\beta} \in \mathbb{R}^{d+1}$ vector of coefficients, $f_{\boldsymbol{\beta}}$ a real-valued function of $\mathbf{Z}_i$, and ε_i a suitable error term independent from $\mathbf{Z}_i$. The core of the conventional approach is utilizing the conditional distribution of U_i given $\mathbf{Z}_i$ for estimation. Bergeron et al. [2] showed that, with left truncation, the conditional likelihood $\mathcal{L}_I$ yields biased estimation because it ignores the information carried by the selection-biased covariates. Moreover, they established that, in contrast, grounding the analysis in the joint distribution of the covariates and response incorporates

this information into the analysis and, consequently, produces superior estimation. Let $\mathcal{L}_J$ denote the joint likelihood. Note that $\mathcal{L}_J(\boldsymbol{\theta}) = \mathcal{L}_I(\boldsymbol{\theta})\, p_{\mathbf{Z}^*}(z)$, with $p_{\mathbf{Z}^*}(z)$ representing the distribution of the biased covariate $\mathbf{Z}^*$. Following the same view as Bergeron et al., we study the impact of the choice of the likelihood function on variable selection and classification.

4 Variable selection

Suppose that in equation (1), $f_{\boldsymbol{\beta}}(\mathbf{Z}_i) = \beta_0 + \beta_1 Z_1 + \cdots + \beta_d Z_d$. Although, training data $\tilde{S}_c$ contains d-dimensional features $\mathbf{Z}_i^*$, not all of them are necessarily related to U_i. In other words, if $\boldsymbol{\beta}^0 = (\beta_1^0, \ldots, \beta_d^0)$ is the true underlying regression coefficient, then there may exist some j's with $1 \leq j \leq d$ such that $\beta_j^0 = 0$. Variable selection in this setting involves finding truly non-zero entries of the regression coefficient $\boldsymbol{\beta}$ in (1) using the train data $\tilde{S}_c$. Let an arbitrary model M be denoted by the set of indices of non-zero entries of the corresponding regression coefficient, i.e., for instance, $M = \{j_1, j_2\}$ is the model corresponding to the family of all $\boldsymbol{\beta}$'s with their only β_{j_1}th and β_{j_2}th entries being non-zero, where $1 \leq j_1, j_2 \leq d$. (Note that such $\boldsymbol{\beta}$ is not unique.) Let M^0 denote the true model, i.e., the model corresponding to $\boldsymbol{\beta}^0$, and $\mathcal{M}$ be the set of all candidate models. $\mathcal{M}$ may or may not include M^0. Following [8], we define

- $\mathcal{M}_o := \{M \in \mathcal{M} : M^0 \subseteq M\}$, called the set of *correct* models, and
- $\mathcal{M}_u := \mathcal{M} \setminus \mathcal{M}_o$, set of *incorrect* models.

Correct models might be inefficient because of their superfluous complexity. The optimal candidate model is the least complex model in $\mathcal{M}_o$, denoted by $M^\star$. Selecting a model from $\mathcal{M}_u$ means missing at least one of the true predictors and, hence, must be avoided. Particularly, as far as revealing the true risk factors is the main interest, it is ultimately desirable for a learning algorithm whose purpose is discovering the true non-zero coefficients not to choose an underfitted model, i.e., from $\mathcal{M}_u$. In the literature numerous likelihood-based criteria has been introduced for model selection (which is, in the present setting, equivalent to variable selection). Examples include Akaike Information Criterion (AIC) [1], Bayesian Information Criterion (BIC) [7], etc, among others. It can be shown that, under length-biased and right-censored training data, employing $\mathcal{L}_J$ as the baseline likelihood function for variable selection is more efficient compared to its counterpart $\mathcal{L}_I$. Roughly, a selection criterion based on $\mathcal{L}_J$ takes less samples to find $M^\star$ in comparison to it being based on $\mathcal{L}_I$. In addition to theoretical justifications, simulation studies, presented in section Simulation Study, also supported this suggestion.

5 Classification

Recently, adopting machine learning techniques has attracted huge attention amongst researchers and policy makers in health care related domains. To a considerable extent, this has been attributed to the swiftly increasing availability of patient records through electronic health records data (EHR). EHR provide access to a large amount of rich data extracted from clinical and administrative data bases. An important question in patient care management is to predict the risk of experiencing a certain outcome, e.g., recurring a health condition, within a particular time frame, say 1 year. However, due to the specific properties of EHR, including length bias and censoring, most of well-known learning techniques cannot be applied naively. Several ad hoc approaches have been tried previously to adapt some machine learning techniques to EHR data but these methods either involve even further loss of information, e.g., by ignoring censored objects, or require the data to be tweaked unnaturally (see [10]). On the other hand, there have been several successful treatments of right-censored data using, for instance, support vector machines, decision trees, and random forests (see, e.g., [3, 4, 5, 6, 9], among others). What is mostly missing in the literature is difficulties induced by left-truncation. This is another problem that we try to address appropriately when it comes to the aforementioned classification question. That is, in presence of length bias how one may correctly model the occurrence of the terminating event in a certain time interval, especially, we focus on the selection bias imposed to the covariates. This may play an important role in methods that rely essentially on the characteristics of the covariate or input space such as tree-based methods.

6 Simulation study

Figure 2 demonstrates the results obtained from BIC variable selection, based on $\mathcal{L}_I$ and $\mathcal{L}_J$ (BIC_I, BIC_j), on simulated, length-biased data generated as follows: (1) Fix $\lambda > 0, \boldsymbol{\beta}$, and sample size n; (2) generate $\mathbf{Z} \in \mathcal{Z} := \{0,1\}^2$ according to discrete uniform distribution over $\mathcal{Z}$; (3) generate U such that $U|\mathbf{Z} \sim \text{Exp}(\lambda e^{\mathbf{Z}\boldsymbol{\beta}})$; and (4) truncate and censor the data based on pre-decided criteria.

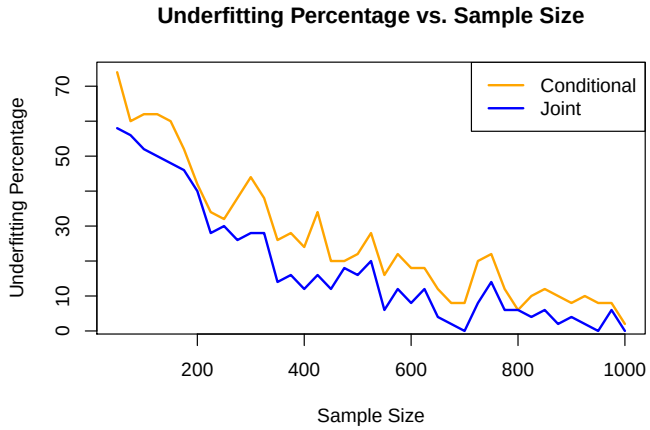

Figure 2: As n increases, underfitting percentage for both BIC_I and BIC_j approaches 0 (consistency)

Size n has been gradually increased from $n = 50$ to 1000 by steps of 25. For each size, 50 different datasets were randomly generated. Figure 2 compares the estimated probability of selecting an underfitted model, i.e., a model containing either Z_1 or Z_2, applying BIC_I and BIC_J. As depicted, BIC_J constantly exhibits a slightly superior performance.

References

[1] H. Akaike. A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6):716–723, 1974.

[2] Pierre-Jérôme Bergeron, Masoud Asgharian, and David B Wolfson. Covariate bias induced by length-biased sampling of failure times. Journal of the American Statistical Association, 103(482):737–742, 2008.

[3] Yair Goldberg and Michael R. Kosorok. Support vector regression for right censored data. Electron. J. Statist., 11(1):532–569, 2017.

[4] Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. Ann. Appl. Stat., 2(3):841–860, 09 2008.

[5] F. M. Khan and V. B. Zubek. Support vector regression for censored data (svrc): A novel tool for survival analysis. In 2008 Eighth IEEE International Conference on Data Mining, pages 863–868, Dec 2008.

[6] Margaux Luck, Tristan Sylvain, Joseph Paul Cohen, Héloïse Cardinal, Andrea Lodi, and Yoshua Bengio. Learning to rank for censored survival data. CoRR, abs/1806.01984, 2018.

[7] Gideon Schwarz. Estimating the dimension of a model. Ann. Statist., 6(2):461–464, 03 1978.

[8] Jun Shao. Linear model selection by cross-validation. Journal of the American Statistical Association, 88(422):486–494, 1993.

[9] Pannagadatta Shivaswamy, Wei Chu, and Martin Jansche. A support vector approach to censored targets. pages 655–660, 11 2007.

[10] David M. Vock, Julian Wolfson, Sunayan Bandyopadhyay, Gediminas Adomavicius, Paul E. Johnson, Gabriela Vazquez-Benitez, and Patrick J. O'Connor. Adapting machine learning techniques to censored time-to-event health record data: A general-purpose approach using inverse probability of censoring weighting. Journal of Biomedical Informatics, 61:119–131, June 2016.