PolyPublie
Polytechnique Montréal

POLYTECHNIQUE MONTRÉAL
UNIVERSITÉ D'INGÉNIERIE

Titre: Title:	Weakly-Supervised Learning from Incomplete Data
Auteur: Author:	Damoon Robatian
Date:	2021
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Robatian, D. (2021). Weakly-Supervised Learning from Incomplete Data [Thèse de doctorat, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/6622/

Document en libre accès dans PolyPublie
Open Access document in PolyPublie

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/6622/
Directeurs de recherche: Advisors:	François Soumis, & Masoud Asgharian
Programme: Program:	Mathématiques

Ce fichier a été téléchargé à partir de PolyPublie, le dépôt institutionnel de Polytechnique Montréal
This file has been downloaded from PolyPublie, the institutional repository of Polytechnique Montréal

https://publications.polymtl.ca

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Weakly-Supervised Learning from Incomplete Data

DAMOON ROBATIAN
Département de mathématiques et de génie industriel

Thèse présentée en vue de l'obtention du diplôme de *Philosophiæ Doctor*
Mathématiques

Mai 2021

© Damoon Robatian, 2021.

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Cette thèse intitulée :

Weakly-Supervised Learning from Incomplete Data

présentée par **Damoon ROBATIAN**
en vue de l'obtention du diplôme de *Philosophiæ Doctor*
a été dûment acceptée par le jury d'examen constitué de :

Luc ADJENGUE, président
François SOUMIS, membre et directeur de recherche
Masoud ASGHARIAN, membre et codirecteur de recherche
Camélia DADOUCHI, membre
Yi YANG, membre externe

DEDICATION

We have not succeeded in answering all our problems. The answers we have found only serve to raise a whole set of new questions. In some ways we feel we are as confused as ever, but we believe we are confused on a higher level and about more important things.

Posted outside the mathematics reading room,
TromsøUniversity

ACKNOWLEDGEMENTS

Firstly, I would like to express my extreme gratitude to my supervisors, Masoud Asgharian and François Soumis, for their consistent support and invaluable guidance throughout this project. Without François' unconditional support at every moment, and Masoud's dedicated help and deep insight into the subject matter, this research would not have been possible.

Further, I would like to extend my special thanks to the members of the jury, Luc Adjengue, Yi Yang, and Camélia Dadouchi, for their priceless time and thoughtful comments and recommendations on this dissertation, which, literally, gave a new life to it.

I cannot forget to express my biggest thanks to the late Serhiî Kolyada, for all he taught me during my previous PhD and Master's degree. No doubt that he played a crucial role in forming my today's passion for research and knowledge discovery.

My family's generous and compassionate support, which was right there whenever I needed it, made this journey, considerably, less agonizing. So, I wish to show my deepest appreciation to them and thank them for everything they did for me. I also, sincerely thank Iryna, for putting up with my endless and timeless hours of working, stresses and moans during the past years. It was, undoubtedly, not less draining and exhausting for her than for me.

Finally, I wish to acknowledge the role of the following people who helped me, in different ways, finalize this research: Vahid Partovi Nia, for providing me with the PhD position and for introducing me to François and Masoud; Farhad Shokoohi, for his help with the simulation study in Chapter 6; Louis-Marc Mercier, for helping me with the French translation of the abstract; my colleagues at the GERAD and Polytechnique offices, for being so amazing and making such an inspiring atmosphere; and the administration of AÉCSP, for granting me the Fonds d'urgence de l'AÉCSP in fall 2020.

RÉSUMÉ

Le paradigme paramétrique de l'inférence statistique a été principalement restructuré au début du XXe siècle. Bien que, pendant plusieurs décennies, cette approche classique de l'inférence ait continué à conserver sa prédominance en tant que principal motif d'inférence accepté, le siècle précédent a également vu l'arrivée d'autres approches favorables à l'inférence statistique. Les techniques non paramétriques, l'analyse exploratoire des données et la théorie de l'apprentissage statistique sont des exemples de ces paradigmes alternatifs. L'adoption de ces nouvelles méthodologies, en plus des avantages offerts par les machines classiques, a enrichi les sciences, où l'analyse des données empiriques est préoccupante.

Cependant, l'adoption des nouvelles méthodologies s'est faite à son propre rythme dans différents domaines. En particulier, dans les domaines où les données communément rencontrées présentent des propriétés atypiques, cette transition a souvent été retardée. La thèse présentée ici étudie et établit certaines étapes essentielles vers la réalisation de l'application du paradigme inférentiel de l'apprentissage statistique ou de la théorie Vapnik-Chervonenkis (VC) à l'analyse de données incomplètes; plus précisément, les données biaisées et censurées sont au cœur de notre intérêt. Ce cadre d'apprentissage sera appelé faiblement supervisé, tout au long de ce travail. En outre, nous étudions les capacités d'apprentissage des réseaux de neurones dits de cartographie, selon les normes fournies par la théorie de l'apprentissage statistique.

La théorie de l'apprentissage statistique, à l'heure actuelle, est l'une des branches les plus matures de la science des données moderne et héberge une riche offre de techniques mathématiquement approuvées pour résoudre les problèmes de données. Comme son homologue classique, les techniques fournies par la théorie de l'apprentissage statistique peuvent alimenter les méthodologies d'analyse de données, en général, et l'analyse de données incomplètes, ce qui est notre intérêt, en particulier. Le certain type de données considéré, ici, est fréquemment rencontré dans l'analyse du temps jusqu'à l'événement ou de la survie, où l'approche paramétrique classique de l'inférence statistique est toujours courante. Cela suggère que l'analyse des données incomplètes pourrait, de manière significative, bénéficier du potentiel des nouvelles méthodes offertes par l'apprentissage statistique.

Bien que certains cadres non classiques aient déjà fait leur chemin dans l'analyse des données de survie, les études fondamentales connexes franchissent encore des étapes rudimentaires. La majorité des études existantes portent sur les performances pratiques de certains algorithmes, tels que les méthodes d'ensemble, sur des ensembles de données de survie concrets. Cela laisse

un certain nombre de questions fondamentales, visant l'applicabilité globale des méthodes d'apprentissage à l'analyse de données incomplètes, sans réponse. Un exemple de telles questions est les conditions nécessaires et suffisantes pour la fiabilité d'une machine pour apprendre à partir de données biaisées et censurées. Combler ces lacunes était la principale motivation derrière la thèse présentée ici. Plus précisément, le paramétrage particulier du biais d'échantillonnage simultané et de la censure semble totalement absent de la littérature sur l'apprentissage statistique.

Pour au moins deux raisons, il est important d'étudier l'intégration des méthodes d'apprentissage statistique dans l'analyse de données incomplètes: (i) La théorie de l'apprentissage statistique offre une flexibilité inférentielle étendue, ce qui conduit à couvrir un plus large éventail de situations dans des problèmes du monde réel; et (ii) il pourrait utiliser la capacité de calcul des ordinateurs modernes pour résoudre des problèmes complexes ou complexes d'analyse de données. De plus, à partir de maintenant, certaines techniques d'apprentissage se sont déjà révélées très prometteuses dans la pratique. La puissance accrue offerte par ces méthodes est ce qui justifie la pertinence de la présente recherche.Nous étudions ici les sujets suivants, dans le cadre de données biaisées et censurées: (i) l'apprentissage de la fonction de distribution, (ii) le problème de minimisation des risques et sa cohérence, (iii) l'apprentissage de la fonction de régression, (iv) la sélection de la variable de régression basée sur l'estimation du maximum de vraisemblance, (v) l'application de la cartographie des réseaux de neurones pour résoudre le problème d'apprentissage, et (vi) certains problèmes importants et ouverts ainsi que quelques défis à considérer dans les études futures.

Les trois premiers problèmes font partie des problèmes les plus fondamentaux de la théorie de l'apprentissage statistique et sont résolus ici, avec succès, pour des données biaisées et censurées. En particulier, nous dérivons les mesures de probabilité empiriques appropriées, définies en termes de données biaisées et censurées, qui peuvent estimer de manière cohérente les mesures de probabilité réelles sous-jacentes. Il est illustré comment les résultats peuvent être appliqués davantage pour minimiser le risque fonctionnel. De plus, une méthode de régression par noyau pour une estimation correcte de la fonction de régression, en présence de biais et de censure, est proposée. En outre, les conséquences de certaines approches naïves du problème sont indiquées.

Dans le cadre du quatrième problème, nous considérons deux méthodes de sélection de variables basées sur la vraisemblance, appelées approches conditionnelle et conjointe. La première est, en fait, basée sur l'approche conventionnelle de l'analyse de régression, c'est-à-dire conditionnant la vraisemblance de la réponse sur les covariables. Elle est dite conventionnelle car elle est basée sur la définition de la fonction de régression et est la méthode, normale-

ment, utilisée dans les problèmes de régression. En revanche, nous proposons une deuxième approche qui utilise la vraisemblance conjointe des covariables et la réponse pour sélectionner les variables. Cette approche a été créée à l'origine pour l'estimation des paramètres mais nous étendons son application au problème de la sélection des variables. Nous pensons que cette dernière approche pourrait être supérieure à la première en termes de sélection du sous-ensemble correct de caractéristiques influentes. Certaines propriétés mathématiques des deux méthodes, qui seraient responsables de cette supériorité, sont dérivées et discutées. Enfin, une brève étude de simulation établit une comparaison entre les deux approches, dont le résultat soutient l'hypothèse de la suprématie de l'approche inconditionnelle sur l'autre. Cependant, fournir une preuve mathématique complète, en faveur ou contre cette hypothèse, nécessite une enquête plus approfondie.

Ensuite, nous fournissons une étude complète des réseaux de neurones dits de cartographie et de leur capacité à résoudre le problème principal de l'apprentissage statistique. Nous remontons les racines mathématiques justifiant les capacités d'estimation des réseaux il y a plus d'un siècle. Divers problèmes mathématiques connexes, tels que le problème de solvabilité algébrique et le théorème de représentation de Kolmogorov-Arnold, sont introduits et leur relation avec les réseaux de neurones cartographiques est examinée. Nous montrons comment la distinction entre l'approximation et la représentation explique la capacité de ces réseaux en estimation de fonction. Pour conclure, nous discutons de la pertinence des réseaux de neurones dans le cadre de la théorie de l'apprentissage statistique. Plus précisément, nous discutons des raisons pour lesquelles les réseaux de neurones ne sont pas capables de résoudre complètement le problème d'apprentissage, selon les principes d'apprentissage statistique d'un apprentissage fiable.

Enfin, quelques défis ouverts, y compris le problème de classification, la détection de la direction de la dépendance et l'apprentissage de la dimension intrinsèque des données, dans le contexte de données biaisées et censurées, sont introduites et, rapidement, discutées.

Comme mentionné précédemment, certains des problèmes résolus ici, tels que l'estimation de la fonction de distribution et le problème de minimisation des risques, sont d'une importance cruciale en théorie de l'apprentissage statistique. La raison, comme expliqué au chapitre 3, est que la résolution des principales formes génériques des problèmes d'apprentissage supervisé, c'est-à-dire l'estimation de la densité, la régression et la classification ou la reconnaissance de formes, se résume à résoudre les problèmes de minimisation des risques et d'estimation de la distribution.

ABSTRACT

The parametric paradigm of statistical inference was mostly systematized in the early twentieth century. Although for several decades this classical approach to inference continued to preserve its dominance as the main accepted ground for inference, the previous century has witnessed the arrival of other propitious approaches to statistical inference too. The non-parametric techniques, exploratory data analysis, and statistical learning theory are all examples of these alternative paradigms. Adopting these new methodologies, in addition to the benefits the classical machinery offers, has enriched sciences, where analysis of empirical data is of concern.

However, embracing the new methodologies has occurred at its own pace in different areas. Particularly, in domains, where commonly encountered data exhibit some atypical properties, this transition has often been delayed. The current thesis studies and establishes some essential steps towards realizing the application of the inferential paradigm of statistical learning or Vapnik-Chervonenkis (VC) theory to the analysis of incomplete data; specifically, the data that are biased and censored, are at the core of our interest. This setting of learning will be called weakly-supervised, throughout this work. In addition, we investigate the learning capabilities of the so-called mapping neural networks, according to the standards provided by statistical learning theory.

Statistical learning theory, by now, is one of the maturest branches of modern data science and hosts a rich supply of mathematically approved techniques for solving data problems. Like its classical counterpart, the techniques provided by statistical learning theory can empower data analysis methodologies, in general, and analysis of incomplete data, which is our interest, in particular. The certain type of data considered, here, is frequently encountered in time-to-event or survival analysis, where the classical parametric approach to statistical inference is still the mainstream. This suggests that incomplete-data analysis might, significantly, benefit from the potential of the new methods offered by statistical learning.

Although some non-classical frameworks have already made their way into the analysis of survival data, related foundational studies are still passing rudimentary stages. The majority of the existing studies deal with the practical performance of certain algorithms, such as ensemble methods, to concrete survival datasets. This leaves a number of fundamental questions, targeting the global applicability of the learning methods to the analysis of incomplete data, unanswered. An example of such questions is the necessary and sufficient conditions for the reliability of a machine for learning from biased and censored data. Filling such gaps

was the primary motivation that triggered the present research. Specifically, the particular setting of the simultaneous sampling bias and censoring seems to be, completely, lacking in the statistical learning literature.

For at least two reasons, it is important to investigate the integration of the statistical learning methods into the analysis of incomplete data: (i) Statistical learning theory provides extended inferential flexibility, which leads to covering a wider range of situations in real–world problems; and (ii) it might employ the computational capability of modern computers to solve complex or computationally heavy problems of data analysis. In addition, as of now, some learning techniques have already proved to be very promising in practice. The increased power offered by these methods is what justifies the relevance of the present research.

Here, we study the following topics, in the context of biased, and censored data: (i) learning the distribution function, (ii) risk minimization problem and its consistency, (iii) learning the regression function, (iv) regression variable selection based on maximum likelihood estimation, (v) application of the mapping neural networks to solve the learning problem, and (vi) some important, open problems as well as a few challenges to be considered in future studies.

The first three problems are amongst the most fundamental problems of statistical learning theory and are settled here, successfully, for biased and censored data. In particular, we derive the appropriate empirical probability measures, defined in terms of biased and censored data, that can consistently estimate the underlying actual probability measures. It is illustrated how the results can be further applied to minimize the risk functional. Also, a kernel regression method for a proper estimation of the regression function, in the presence of bias and censoring, is proposed. In addition, the consequences of some naïve approaches to the problem are indicated.

In connection with the fourth problem, we consider two likelihood-based variable selection methods, referred to as the conditional and joint approaches. The first one is, in fact, based on the conventional approach to regression analysis, i.e., conditioning the likelihood of the response on the covariates. It is called conventional because it is based on the definition of the regression function and is the method, normally, used in regression problems. In contrast, we propose a second approach that employs the joint likelihood of the covariates and the response for selecting variables. This approach was originally created for parameter estimation but we extend its application to the problem of variable selection. We speculate that the latter approach might be superior to the former one in terms of selecting the correct subset of influential features. Some mathematical properties of both methods, which are believed to be responsible for this superiority, are derived and discussed. Finally, a brief

simulation study draws a comparison between the two approaches, whose outcome supports the hypothesis of the supremacy of the unconditional approach over the other one. However, providing a complete mathematical proof, in favour of or against this hypothesis, requires further investigation.

Next, we provide a comprehensive investigation of the so-called mapping neural networks and their capability of solving the main problem of statistical learning. We trace back the mathematical roots justifying the estimation abilities of the networks to more than a century ago. Various related mathematical problems, such as the algebraic solvability problem and the Kolmogorov-Arnold representation theorem, are introduced and their relation with the mapping neural networks are scrutinized. We show how the distinction between the approximation and representation explains the capacity of these networks in function estimation. To conclude, we discuss the relevance of the neural networks inside the framework of statistical learning theory. Specifically, we discuss why neural networks are not able to, completely, solve the learning problem, according to the statistical learning principles of reliable learning.

Finally, a few open challenges, including the classification problem, detection of the dependency direction, and learning the intrinsic dimension of data, in the context of biased and censored data, are introduced and, swiftly, discussed.

As mentioned earlier, some of the problems solved here, such as estimation of the distribution function and the risk minimization problem, are of crucial importance in statistical learning theory. The reason, as explained in Chapter 3, is that solving the main generic forms of the supervised learning problems, i.e., density estimation, regression, and classification or pattern recognition, boil down to solving the risk minimization and the distribution estimation problems.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ACRONYMS

AFT accelerated failure time. 62, 64, 65, 90

AIC Akaike information criterion. 12, 24, 92, 95, 96, 108

BIC Bayesian information criterion. 12, 92, 95, 96, 108

CDF cumulative distribution function. 8, 51, 67, 68, 74

CPH Cox proportional-hazards. 26, 62–64

DA discriminant analysis. 20

EDA exploratory data analysis. viii, 22

EHR electronic health records. 25, 138

ERM empirical risk minimization. 15, 36, 38, 39, 41–43, 137

FIC focused information criterion. 12

GLM generalized linear models. 23

HQC Hannan–Quinn information criterion. 12

i.i.d. independent and identically distributed. 8, 36, 37, 46–48, 66–68, 97, 115

LASSO Least Absolute Shrinkage and Selection Operator. 26, 92, 95

LBRC length-biased, right-censored. 3–5, 8, 10, 11, 69

LBRC-C length-biased, right-censored, with covariates. 1, 2, 6, 8–10, 12, 15, 16, 18, 61, 66–70, 72, 77–80, 82, 83, 85, 86, 90–92, 96–98, 103, 105, 137

LLN law of large numbers. 39–42, 46, 48

MLE maximum likelihood estimation. ix, 5, 11, 16, 20–24, 68, 85, 90, 91, 95, 97, 101, 105, 115

NPMLE nonparametric maximum likelihood estimation. 69

PDF probability distribution function. 52

SCAD Smoothly Clipped Absolute Deviation. 92

SRM structural risk minimization. 7, 15, 44–46, 137

SVM support vector machine. 1, 7, 18, 24, 115

VC Vapnik-Chervonenkis. viii, 7, 19, 32, 40, 45, 47, 49, 93

WAIC Watanabe–Akaike information criterion. 12

Nomenclature

$\overset{*}{\mathbf{X}}$	biased covariate vector
$\xrightarrow{a.s.}$	almost sure convergence
$\xrightarrow{p}$	convergence in probability
$\mathcal{D}$	sample data
$\succcurlyeq, \preccurlyeq$	entry-wise inequality of vectors
$\langle ., . \rangle$	inner product
$\tilde{A}$	length-biased current lifetime
$\tilde{R}$	length-biased residual lifetime (survival time)
$\tilde{Y}$	length-biased response variable
$\overline{T}$	left truncation
$\overset{*}{.}$	bias (*not* length bias)
$^\top$	matrix transpose
$\tilde{.}$	length bias
$\wedge$	min, minimum
C	residual censoring time
C'	total (overall) censoring time
Y	response variable

CHAPTER 1 INTRODUCTION

1.1 Initial Motivation

How to make a *statistically reliable inference* about a population of interest, when the available sample data are *incomplete* is what the current thesis studies. The certain type of incompleteness considered, i.e., *sampling bias* and *censoring*, is frequently encountered in observational studies, particularly, in the so-called *follow-up, prevalent-cohort, cross-sectional* study design. However, a comprehensive investigation of the *learning* problem in this specific setting is missing, completely, in the statistical learning literature. This deficiency was the initial motivation that convinced us to undertake this research project. The present study, successfully, and to a considerable extent, fills in some of the existing gaps in the area of learning from length-biased, right-censored, with covariates (LBRC-C) data and paves the way for filling further gaps in future. The detailed contributions of our study are listed in the upcoming sections of this chapter.

1.2 General Context and Relevance

The specific incompleteness being considered consists of two components: (i) A type of sampling bias, called *length bias*, and (ii) *right censoring*, which is a partial loss of information on some of the sampled units. Making statistical inference in presence of length bias and right censoring has a quite long history in several fields including *survival analysis* and *reliability theory*, due to its practical usefulness and popularity. When the response variable of interest in data analysis is of *time-to-event* sort of nature, considering the length bias and censoring is almost inevitable, because of various restrictions the researchers commonly have to face. The classical approaches to time-to-event or survival analysis are predominantly based on the inferential framework of classical statistics, which was formally established in the early twentieth century, almost exclusively, by the works of Ronald A. Fisher (1890–1962) [Vapnik, 1998].

Apart from the invaluable contributions of the classical approach to the development of statistical inference and its applications in other areas of science, including biology and medicine, several other approaches made their way into the world of data analysis as well. Moreover, some of them proved to be quite promising in various applications. For instance, some methods of statistical learning theory, such as the support vector machines (SVMs) and ensemble methods as well as different types of artificial neural networks are among the state-

of-the-art methods of today's data analysis [Horne et al., 2009, Sidey-Gibbons and Sidey-Gibbons, 2019, Matheny et al., 2020]. This raises a natural question: Could incorporating the statistical learning inferential machinery into time-to-event data analysis, possibly, empower the existing inferential framework? In particular, how it affects the learning methodology in the context of LBRC-C data?

There are reasons to believe that the answer to the aforementioned question is positive: First, the *flexibility* offered by more recent *nonparametric* approaches, including that of statistical learning theory, facilitates making an inference in a broader range of problems where, typically, the assumptions of parametric statistics could hardly be met. Second, the development of statistical learning theory, was motivated, to some extent, by the advent of new powerful computers in 1960s. Hence, being computer compatible has always been at the very core of the statistical learning methodology. By consolidating statistical inference with the *computational capacity* of modern computers, statistical learning might make a considerable contribution to the world of data analysis. (See chapters 2 and 3 for more details.)

Nonetheless, statistical learning has only begun to establish its potential in numerous areas, where the classical paradigm of inference is still considered mainstream. For instance, in many fields of biology, medical sciences, psychology, etc. This is not surprising as it is well known that these sciences and statistics have been benefited from each other since a long time ago [Halpin and Stam, 2006]. In fact, some of the most celebrated statistical tools were first motivated by applications in biology and related fields. Examples include testing hypothesis, randomized controlled trials for treatment assessment, and etc [Fisher, 1935].

One of the areas that need to be investigated more, from the statistical learning theory point of view, is undoubtedly the analysis of time-to-event data. This particular type of data is found in a diverse range of disciplines, including those studying the lifespan of both living organisms and other objects. While there is already a considerable amount of research in the context of time-to-event or survival data in the framework of survival analysis, reliability engineering and machine learning, some more specific areas have attracted less attention. Analysis of LBRC-C data is one of them. In particular, when it comes to statistical learning theory, the literature falls short. To the best of our knowledge, there is not even a single account of research that, specifically, studies learning from LBRC-C data. This significant shortage makes our research relevant and inescapable.

The present thesis, hence, is dedicated to investigating multiple aspects of learning from LBRC-C data. Some of the problems considered, particularly, aim at fundamental aspects of learning in the context of interest such as estimating the distribution function or risk

minimization, while others target ultimate applicability of certain tools to solve the learning problem, e.g., investigating the capability of the mapping neural networks in function estimation based on length-biased, right-censored (LBRC) data. The details of the particular problems considered in the present study will be given later in this chapter.

1.3 Background and Importance of Studying LBRC-C Data

Time-to-event is the output of interest in numerous disciplines spanning epidemiology, economics, econometrics, gerontology, and etc [LeClere, 2005, Backman et al., 2011, Asher et al., 2017]. It is defined as the amount of time elapsed from the occurrence of an *initiating event* until that of a second event called a *terminating event*. Both events are pre-defined. For example, the initiating event might be birth, the 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, 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 other measurements for that individual.

Naturally, any data-driven inference depends on the 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, also, several concerns worth considering. One of the most crucial factors, ignoring which may cause serious issues, is data *incompleteness* [Nakagawa and Freckleton, 2008]. We usually have to deal with a compound incompleteness consisting of multi-aspect distortion and information loss. The following paragraph elaborates more on this issue.

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 [Wang et al., 1993, Asgharian and Wolfson, 2005, Bergeron et al., 2008]. 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*. 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. This leads to a phenomenon referred to as *left truncation*. Due to the presence of right censoring, learning from such data for prediction and generalization falls into one of the subcategories of well-known *supervised* learning. We call this subcategory *weakly-supervised* learning as the information on the right censored responses is only partially available. As such, a prevalent cohort comprises a non-random sample that is not representative of the target incident population.

Left truncation and right censoring has been extensively studied in survival analysis; particularly, right censoring, because of its prevalence. The consequences of failure to take the left truncation into account has been discussed by Wolfson et al. [2001]. In particular, it was illustrated how this failure, almost surely, results in an overestimation of the survival time in patients with dementia. For more details on left truncation and right censoring, see Lagakos et al. [1988], Leung et al. [1997], Barrajon and Barrajon [2019], Lagakos [1979], Prinja et al. [2010], among others.

Length bias, however, which is a special case of left truncation, has been studied much less. Length bias occurs when the chance of being selected into the sample is proportional to the survival time. For a general survey of methods for data analysis with length bias and their applications, one may refer to Arratia et al. [2019]. An important feature of length biased data, when covariates are also collected for each subject, is an additional layer of bias introduced to the sampling distribution of the covariates. This specific aspect of length bias had been completely ignored in the literature until recently when Bergeron [2006] and Bergeron et al. [2008] addressed the problem.

As mentioned earlier, the current thesis pioneers studying the problem of learning from LBRC data in presence of covariates, in the statistical-learning-related literature. Studying the learning characteristics of LBRC-data analysis seems inevitable according to the frequent application of the prevalent-cohort cross-sectional sampling design with follow-up in time-to-event and survival analysis. For instance, Huang and Wang [1995], Wang [1991], Wang et al. [1993], all consider this selection setting for conducting their research.

Surprisingly, some of the problems considered here, such as the regression function estimation, discussed in Chapter 4, as well as the variable selection problem, studied in Chapter 5, had no mention even in the survival analysis literature. Nonetheless, in the past couple of years and due to the increasing accessibility of recording data digitally, sampling bias and incompleteness became more visible and have sparsely attracted the attention of some researchers active in areas close to learning theory. Examples are Luck et al. [2018], Laforgue and Clémençon [2019]. Note that none of the existing works, including Luck et al. [2018], Laforgue and Clémençon [2019], considered learning from LBRC data.

1.4 Inference in Statistical Learning Theory

The appearance of statistical learning theory and its development was a response to certain limitations of the classical approach to statistical inference. By "classical" we mainly refer to the *parametric* paradigm of statistics, whose first systematic application to solving data problems was thanks to Ronald Fisher in 1920s. The major ingredients of the presently well-known parametric statistics had been around long before that time, but it was him who, in addition to systematizing its application for the first time, promoted and popularized it on a global scale. The parametric approach remained the main, or probably the only, widely accepted paradigm for several decades [Vapnik, 1998].

Perhaps the most essential elements of the parametric approach are (i) parametric families of distributions such as the exponential family, and (ii) the powerful maximum likelihood estimation (MLE) as its primary inferential engine. Both of these elements, despite being extremely powerful, impose some restrictions that limit the applicability of the parametric framework in many situations [Vapnik, 1998, 1995]. For example, the theoretical validity of inference based on parametric distributions highly depends on strong assumptions that need to be satisfied a priori. The problem is that many of these constraints are hardly met in the world of real problems. The MLE principle, also, suffers from several issues that make it not always the first choice of preference. These problems will be discussed in more details in Chapter 3.

Moreover, the arrival of the first powerful computers in 1960s made it possible to put the classical methodology to the test, particularly, in situations that previously were just beyond the reach of traditional methods due to their computational complexity [Vapnik, 1998, James et al., 2013]. This further reveals the necessity to rethink some aspects of the mainstream statistical inference. All of these events together led to several new directions in the analysis of data, including statistical learning theory.

The details of statistical inference in the framework of statistical learning is discussed in Chapter 3. Hence, here we skip the in-depth discussion of the theory of statistical learning. Nonetheless, let us briefly introduce the main building blocks of statistical learning theory, in order to motivate its use in the context of survival analysis with LBRC-C data [Vapnik, 1992, 1995, 1998].

1. Statistical learning theory establishes the main problem of learning, i.e., the problem of *statistical inference*, as a problem of *function estimation*.

2. Function estimation involves *choosing a function* from a set of pre-determined *admissible* functions (also called the *hypothesis space*), provided a *limited (finite)* amount of empirical data.

3. In choosing the set of admissible functions and then the best function among them, two criteria are always considered:

   (a) Minimum *risk* or maximum *utility* that could be achieved by applying a specific function among the set of admissible functions, in order to make inference inside the set of given empirical examples. This is usually called the *fit* of a model, i.e., how well the specific function that has been chosen fits the data;

   (b) The function's ability to generalize beyond the given set of example data. That is, the ability to achieve a small risk if being applied to new data. This is what we call *generalization ability*.

4. The general quality of the learning process is measured based on the well-known statistical criterion of *consistency*. The *qualitative* part of the theory provides the *necessary* and *sufficient* conditions for the consistency of a learning processes.

5. The general *quantitative* part of the theory includes *bounds of the rates of convergence*, which can be regarded as the rate of generalization of these learning processes.

6. Principles for estimating functions from a *small* collection of data, based on the developed theory.

7. Concrete methods of function estimation and their application to solving real-life problems. These methods must satisfy the previous points.

The first three items above comprise the methodological backbone of statistical learning theory. From the theoretical point of view, the goodness-of-fit and the generalizability of an

individual candidate function, as explained in Chapter 3, might be measured by averaging the risk of applying that function for making an inference, which can be either description or prediction, evaluated on all possible values of data. However, to be able to do so, one needs complete information about the distribution of data. In reality, this information is not available and instead a sample of data is provided, which must be employed to, empirically, estimate the underlying probability measure. This brings us to the main problem of mathematical statistics, i.e., estimation of an actual probability measure by means of an empirical measure constructed base on the available data.

Once and empirical measure is known, one needs to assess both its validity and feasibility for solving the learning problem. At this point, the qualitative and quantitative theories, mentioned in the fourth and fifth items above, come into play. Now, one of the important distinctions of the inferential approach of statistical learning theory with that of the classical one appears: In practice, we are given only a finite amount of sample data. Therefore, it makes sense to verify the viability of the learning process for "small" collections of data. In fact, this is the intersection of the qualitative and the quantitative theories of statistical learning since it provides the relation between the quality of learning, which is, roughly, the difference between the expected and empirical risks, in terms of sample size and the *capacity* of the set of admissible functions.

The capacity of a set of functions, called the *Vapnik-Chervonenkis dimension (VC dimension)*, plays a crucial role in statistical learning theory. It represents the complexity of the learning problem that can be solved by that particular set of functions. As we will see later, this is a major factor in both determining the "right" set of hypotheses as well as the rate of convergence of the empirical risk to the expected one [Vapnik and Chervonenkis, 1974a,b, Vapnik, 1995]. What makes this procedure different from the classical approach is the fact that the potential admissible functions are accepted into the hypothesis space taking both the sample size and the complexity of the problem at hand into account. This is in contrast with constructing the hypothesis space based on the functions' form. Recall that in linear regression, e.g., the hypothesis space consists of all linear-in-parameter functions, with complete disregard for the complexity of the problem and functions' ability to solve the regression problem. The risk minimization task carried out over such a hypothesis space, determined as explained, is called the *structural risk minimization (SRM)* [Vapnik and Chervonenkis, 1974a,b, Vapnik, 1995, 1998].

Finally, statistical learning theory provides concrete methods of function estimation, i.e., solving the main problem of learning, that satisfy the aforementioned principles. SVMs are, perhaps, the most prominent example of such methods.

Now, it is important to note that the entire theory of statistical learning has been originally developed for the situation where the provided data are *representative*. Since our main objective is to focus on learning from biased and censored data with covariates, it is necessary to consider all of the fundamental elements of statistical learning theory, carefully, in the new context of interest and establish the infrastructure once more but with respect to the setting we are interested in. This leads us to the particular list of problems to be addressed, in order to build a coherent framework based on the principles of statistical learning theory for learning from LBRC-C data, including the case where covariates must also be added to the analysis. The next section is devoted to these problems.

1.5 Research Problems and their Solutions

In what follows, we discuss, swiftly, the main problems addressed in the present work. Detailed representation of each problem is postponed to the corresponding chapter. Besides, we introduce the results obtained, briefly.

1.5.1 Learning the Distribution Function from LBRC-C Data

As mentioned in the previous section, one important problem in both statistics and statistical learning theory is estimating the probability measure from empirical data. More precisely, assume that there is a probability space (Ω, Σ, P) that describes the distributional structure of the stochastic phenomenon of interest, with $\mathbf{Z} : \Omega \to \mathbb{R}$ being the corresponding random variable. The problem involves estimating the *unknown* probability measure P based on a given set of independent and identically distributed (i.i.d.) realization of $\mathbf{Z}$, say $\mathcal{D} = \{\mathbf{z}_i : i = 1, 2, \ldots, n\}$. In fact, one should estimate $P(A)$, for any measurable set $A \in \Sigma$. If $\mathcal{D}$ is representative of the target population, then, $P(A)$ is usually estimated by the empirical measure $\hat{P}_n(A)$ defined as

$$\hat{P}_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_A(\mathbf{z}_i),$$

where $\mathbb{1}_A$ is the indicator function of A. A special case of this problem is learning the distribution function, i.e., the cumulative distribution function (CDF), defined on a certain subset of the σ-algebra Σ. Note that this problem is defined in detail in Chapter 4.

For the case of representative data, this problem was completely solved and its asymptotic properties were also established by the well-known multidimensional Glivenko-Cantelli theorem. Also, Asgharian et al. [2002], Asgharian and Wolfson [2005] had, thoroughly, established the problem when data are LBRC but without covariates.

However, estimating the distribution function from LBRC-C data was still an open problem, prior to our work. Therefore, the first problem to be addressed in our research is the following:

P1. *Given the measurable space* (Ω, Σ) *and a set of length-biased and right-censored realizations,* $\mathcal{D}$*, of the random variable (vector)* $\mathbf{Z}$*, defined on* Ω*, estimate the distribution function* $F_{\mathbf{Z}}$*.*

Problem P1 is discussed in section 4.1. In particular, we prove that there are empirical measures that can reliably, in the sense of almost sure consistency, estimate the distribution function base on LBRC-C data. Due to significant differences, we discuss the case where data are purely biased and where right censoring is also present in data, separately. Clearly, in both cases, we assume that the sample data contain a vector of covariates associated with each subject.

1.5.2 Learning the Regression Function from LBRC-C Data

One of the principal problems in the area of supervised learning is the regression problem. In fact, many authors divide the problems of supervised learning into two main subcategories of regression and classification or pattern recognition problems. This division is roughly based on the nature of the response (output) variable. That is, regression refers to the situations where the response variable of interest is a continuous random variable, while classification or pattern recognition involves a discrete response variable.

The regression problem is defined as learning the *regression function* form a set of sample data, including both the covariate and the response values for each sampled subject. That is, estimating the function

$$\mathbb{E}_{Y|\mathbf{X}=\mathbf{x}}(Y) = \int y \, \mathrm{d}F_{Y|\mathbf{X}=\mathbf{x}}(y), \qquad \forall \mathbf{x} \in \mathbb{R}^d;$$

where $\mathbf{X}$ is a vector of covariates and Y is the response. Note that this is the definition of the regression function and can be found in any textbook of mathematical statistics [Casella and Berger, 2002, Vapnik, 1998]. In our setting, the given response values are subject to length bias and right censoring. Two frequently applied approaches to the regression problem can be stated as follows: (i) The regression function $\mathbb{E}_{Y|\mathbf{X}=\mathbf{x}}(Y)$ is assumed to be an *explicit* parametric function of $\mathbf{x}$, say $r_{\boldsymbol{\beta}}(\mathbf{x})$, with $\boldsymbol{\beta} \in \mathcal{B}$, being the so-called regression coefficient. Hence, in this case, the estimation of the regression function boils down to the estimation of the coefficient $\boldsymbol{\beta}$. Note that to be able to solve this problem one requires that the identifiability assumption is held. Commonly, this setting is referred to as *parametric* regression. Nonetheless, we try to avoid this term out of its ambiguity (see Chapter 4) and

call it *explicit*. (ii) No explicit form for the regression function $\mathbb{E}_{Y|\mathbf{X}=\mathbf{x}}(Y)$ is assumed. Let us denote $r(\mathbf{x}) = \mathbb{E}_{Y|\mathbf{X}=\mathbf{x}}(Y)$. Here, the regression estimation requires estimating the value of the regression function $r(\mathbf{x})$, at each point $\mathbf{x}$. Similarly, as in the previous case, we call this the *non-explicit* regression rather than *nonparametric*. Here in this thesis, both cases are of interest but with different motivations.

In the survival analysis literature, the explicit regression problem with length bias and right censoring is very well studied. However, all likelihood-based analyses of the regression problem had failed to take a very subtle but decisive point into account: An additional layer of bias that is induced by the length bias and affects the sampling distribution of the covariates. Bergeron et al. [2008], Bergeron [2006] settled this problem by proposing a new likelihood-based approach which will be referred to as the *joint-likelihood* or simply *joint* approach. This approach properly treats the covariate bias, in contrast to the *conventional* approach, which is based on the likelihood of the response, conditioned on the covariate distribution. Bergeron et al. [2008], illustrated the superiority of the joint approach, compared to the conditional one, in terms of parameter estimation bias and efficiency.

In this thesis, we consider the non-explicit regression problem, i.e., estimating the regression function from LBRC-C data. This leads us to the second major problem being investigated:

P2. *Let* $r(\mathbf{x}) = \mathbb{E}_{Y|\mathbf{X}=\mathbf{x}}(Y)$ *be the non-explicit regression function. Estimate the value of* $r(\mathbf{x})$*, for any covariate vector* $\mathbf{x}$*, given length-biased and right-censored sample data* $\mathcal{D}$*.*

Problem P2 is addressed in section 4.3. As for P1, we investigate, first, the case where data are length biased but no censoring is allowed. Then, separately, we solve P2 for the LBRC case. In both cases, we propose a *kernel* regression solution to the problem, which are, indeed, the generalized counterparts of the well-known Nadaraya-Watson kernel regression. We, also, indicate two different naive applications of the kernel regression to LBRC-C data, which result in incorrect estimates of the regression problem.

1.5.3 Risk Minimization Problem with LBRC-C Data

As discussed earlier, risk minimization is at the heart of the statistical learning methodology for making statistical inference. In fact, the main problem of statistical learning is formulated based on a risk minimization task [Vapnik, 1998]. That is, let $\mathcal{H}$ denote the hypothesis space and $R(h)$ be expected risk associated with any $h \in \mathcal{H}$. Then, one needs to solve the following optimization problem in order to find the optimal functional dependency between the input and output:

$$\inf_{h \in \mathcal{H}} \{R(h)\}.$$

We will see in Chapter 3, R is defined based on the underlying probability measure, which is not known, as briefly explained in subsection 1.5.1. In practice, one uses the *empirical risk functional*, denoted by $\hat{R}_n(h)$, instead of the expected one. $\hat{R}_n(h)$, however, is an empirical process whose consistency has to be established to guarantee the reliability of the learning machine. When data are representative, this result has been provided by Vapnik and Chervonenkis [1968].

Now, the problem of risk minimization and its asymptotic properties in the context of LBRC data was never considered before. This constitutes the next research problem in the present study:

P3. *Given length-biased and right-censored data $\mathcal{D}$ and a suitable loss function, construct an empirical risk functional $\hat{R}_n$ that can, consistently, estimate the expected risk R over the hypothesis space $\mathcal{H}$.*

Note that the covariate-free case is a special case of this more general one, and hence, is implied from it. The same fact applies to the estimation of the distribution function. P3 is studied in section 4.2, where we establish the empirical risk functional for the length biased and for the length-biased, right-censored data, independently. In addition, we provide sufficient conditions for the consistency of the empirical risks.

1.5.4 Regression Variable Selection with LBRC-C Data

Variable (feature) selection is of significant importance from both descriptive and predictive perspectives [Sauer et al., 2013, Chowdhury and Turin, 2020, Genuer et al., 2010a, Meyer et al., 2019]. It is, also, used for a broad variety of purposes, such as detecting the influential factors in a health condition or for initiating preventive measures to avoid adverse outcomes of a treatment. For different applications of variable selection see Ertefaie et al. [2018], Sauer et al. [2013], Chowdhury and Turin [2020], Lu and Petkova [2014], among others. Variable selection in the context of explicit regression, which was defined earlier in subsection 1.5.2, comprises the next problem we address here in this study.

In regression analysis, variable selection involves, systematically, determining the most significant set of covariates out of all available covariates in the sample data. In subsection 1.5.2, we mentioned that, in our investigation, the explicit regression setting would be of interest. In fact, we investigate the variable selection based on the MLE in the explicit regression setting. The entire chapters 5 and 6 are devoted to this problem and a related, short simulation study.

The existing literature on variable selection, in general, is extremely rich. For a review of

variable selection methods one may refer to Desboulets [2018], Heinze et al. [2018], and Sauer et al. [2013]. The problem of variable selection might be considered as a form of *model selection*. There exist various types of criteria for optimal model selection amongst a set of pre-determined candidate models. In many frequently used criteria for model or variable selection, likelihood function constitutes one of the major components of the criteria. This class of variable selection methods is the one we focus on. Examples of likelihood-based criteria are the the *information-based* ones, such as the well-known Akaike information criterion (AIC) and Bayesian information criterion (BIC) [Akaike, 1974, Schwarz, 1978] as well as their modified variants, including Hannan–Quinn information criterion (HQC) [Hannan and Quinn, 1979], Watanabe–Akaike information criterion (WAIC) [Watanabe, 2010, 2013], focused information criterion (FIC) [Claeskens and Hjort, 2003, Hjort and Claeskens, 2003, 2006].

Recall the conditional and joint likelihood-based approaches we have already mentioned in subsection 1.5.2. We explained that the difference between these two methods, in terms of parameter estimation, has been already established by Bergeron et al. [2008], Bergeron [2006]. Clearly, both the conditional and joint approaches can be potentially used in constructing model or variable selection criteria. However, because the joint approach has been created rather recently, it has never been employed for the purpose of model or variable selection. In Chapter 5, first, we propose a family of variable and model selection criteria that are constructed utilizing the joint likelihood function. Note that it is the first time in the literature that the unconditional likelihood has been applied to the problem of selection of variables with LBRC-C data. Then, we study the distinction between the conditional and the joint likelihoods when they are used as the likelihood component of variable selection measures in the context of LBRC-C data. This brings us to the next problem: (Detailed definition is given in Chapter 5.)

P4. *Consider the regression problem*

$$Y = r_{\boldsymbol{\beta}}(\mathbf{X}) + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\beta} \in \mathcal{B},$$

where $Y \in \mathbb{R}$, $\mathbf{X} \in \mathbb{R}^d$, $\mathcal{B} \subseteq \mathbb{R}^{d+1}$, $r_{\boldsymbol{\beta}}$ *is a real-valued linear function of* $\boldsymbol{\beta}$ *and* $\mathbf{X}$, *and* $\boldsymbol{\varepsilon}$ *denotes a suitable error term independent from* $\mathbf{X}$. *Then, if* $\boldsymbol{\beta}^0 = (\boldsymbol{\beta}^0{}_0, \ldots, \boldsymbol{\beta}^0{}_d)$ *is the true regression coefficient, then, investigate and compare the impact of employing the conditional and joint likelihoods on variable selection, given a length-biased and right-censored sample dataset. In particular, which one is more capable of selecting* correct *models? Why?*

Due to the fact that the joint approach uses the thorough information provided by training

data, while its conditional counterpart does not, by ignoring the information contained in the distribution of the covariates, we speculate that the unconditional likelihood should be more efficient in selecting the correct set of variables than the conditional one. This is the core hypothesis that is put to the test in Chapter 5.

In connection with P4, we show that (i) conceptually, the correct choice for variable selection is the joint likelihood and not the conventional (conditional) one; (ii) the two likelihoods exhibit different behaviours in certain respects, which in turn affects the variable selection; (iii) these distinctions might, ultimately, lead to higher chances of selecting incorrect (underfitted) models by the conditional approach. Despite being able to show these divergences between the two approaches, unfortunately, at the moment we cannot provide a rigorous proof for the hypothesis mentioned earlier, however, a simulation study, provided in Chapter 6, seems to be aligned with the superiority of the proposed approach compared to the conventional one.

All the main results obtained in Chapter 5 are *original* contributions of the present thesis.

1.5.5 Learning by Mapping Neural Networks

All questions considered so far aim at providing a global procedure that let one solve a learning problem in a theoretically sound way. Metaphorically, they provide a *road map* that indicates the correct and reliable path to the destination, being the solution to the problem. What still needs to be considered is to find a reliable *vehicle* that can take one to that destination. Assume that we are provided with a list of multiple vehicles that can potentially take us to the destination, however, we need to inspect them and make sure they are capable of completing this task.

Aligned with this metaphor, our purpose of considering the artificial neural networks, in this research, is to find out whether they could be regarded as a vehicle that could be used to reach the destination, and if yes, how well they keep up with the *standards* of statistical learning theory throughout the road. Note that we focus on a specific type of networks, commonly called mapping neural networks. These are perceptron-like multilayer feedforward neural networks with the capability of estimating functions from a finite amount of data. It is worth noting that while our ultimate goal is to study the capability of the mapping neural networks in learning from biased and censored data, here and as a first step, we try to settle the problem in a less specific framework, i.e., learning from complete data.

Regarding the history of neural networks, creation of the first multilayer perceptron by Frank Rosenblatt, in late 1950s, marked a major milestone in the history of both learning theory,

in general, and the neural networks, in particular [Rosenblatt, 1962]. In fact, Rosenblatt's perceptron was the very first actual learning machine that came into existence. Shortly after, the mathematical theory of learning was initiated by Novikoff [1962]. Despite the promising start, artificial neural networks lost their initial appeal soon and began to attract attention only about two decades later, when an already existing gradient-based fitting method [Kelley, 1960, Bryson, 1962] was applied for updating the weights of a neural network, in 1980s [Rumelhart et al., 1986a,b]. This method was called *backpropagation*.

The application of backpropagation brought up the artificial neural networks to the attention of researchers for the second time. This was followed by a series of successful applications of neural networks in particular areas. Most of the studies on artificial neural networks have been focused on the applied aspects of them, as a result of which more fundamental properties were neglected until recent decades. There are still a huge amount of unanswered questions about the mathematical mechanism underlying their success in practice. As a part of this study, we try to take some steps in order to shed some light on the mathematics of the mapping neural networks and their ability to solve the learning problem, especially, with respect to the principles considered in statistical learning theory.

Problem P5, stated below, gives a high-level description of the questions that motivated the material of Chapter 7. We call it "high-level" because the following questions consist of several sub-questions that are addressed, separately, in the dedicated chapter.

P5. *Could mapping neural network be employed to solve the main problem of learning, i.e., the problem of statistical inference? If yes, How compatible are they with the statistical learning methodological standards?*

In order to answer the questions mentioned in P5, we have tracked down the mathematical results that are connected to the problem of function estimation, in some cases, to more than a century ago. Connected to the mapping neural networks, at least three different, but closely related, notions must be distinguished to understand how mapping neural networks might be employed and solve the learning problem. These three notions are (i) function *estimation*, (ii) function *representation*, and (iii) function *approximation*. Estimation refers to the general capability of solving the learning problem and depends on the other two.

As we will see in Chapter 7, while the representation ability of networks is sufficient to guarantee the mapping neural networks' capacity of solving the learning problem, theoretically, when it comes to practice, implementing such networks is extremely hard, if not impossible. On the other hand, for the purpose of learning, having *exact* representation is not necessary and, therefore, out of interest. We will see that a reasonable approximation ability is suf-

ficient for being able to solve the learning problem. In addition, such networks are easy to implement in practice.

Nonetheless, compatibility of the mapping neural networks with the statistical learning ecosystem requires close attention. Recall that the SRM principle comprises one of the key concepts that balance the complexity of the learning machine according to the problem under consideration as well as the sample data. This principle is not satisfied with the mapping neural networks being utilized for solving the learning problem. While, according to statistical learning theory, the learning procedure consists of two *selection* stages, learning by the mapping neural networks completely ignores one of them resulting in solving the learning problem only partially, in the explained sense. Very recently we became aware of a work by Vapnik that points to the same problem [Vapnik, 2019].

1.6 Layout of the Thesis

Chapter 2 is devoted to a discussion on the existing related literature. It is important to notice that the present work consists of components belonging to multiple related but distinct areas: (i) Statistical learning theory, (ii) survival (time-to-event) analysis, and (iii) mapping neural networks. This fact, indeed, gives a strong interdisciplinary character to this work, which means that the topics included might have been motivated and developed, independently and in disparate scientific contexts, nevertheless, in some cases simultaneously. This complicates, particularly, reviewing and providing the pertinent literature in a smooth chronological manner. As a result, we introduce the literature of each subject in a separate section.

In Chapter 3, we introduce the main elements of statistical learning theory and survival analysis, which can be considered as preliminary knowledge for the development of the rest of the thesis. Related to statistical learning theory, particularly we introduce the main problem of statistical learning, the principles of risk minimization, including expected risk minimization, empirical risk minimization (ERM), and structural risk minimization (SRM). We, also, introduce the VC dimension, which plays a pivotal role in the SRM procedure. Finally, we state the *necessary and sufficient* conditions for the *non-trivial consistency* (the definition of which is given in Chapter 3) of learning.

In addition, we provide the main concepts of survival analysis that are required for our investigation in chapters 4 and 5. This includes the details of the setting we are interested in, i.e., LBRC-C data and the prevalent-cohort, cross-sectional sampling design with follow-up.

Chapter 4, discusses problems P1, P2, P3, and related issues. Precisely, we establish the

problem of learning the distribution function, non-explicit regression function and risk minimization from LBRC-C data.

In Chapter 5, problem P4 is settled. Specifically, a new MLE-based variable selection procedure is proposed. The new method is grounded in the joint distribution of the covariates and the response rather than the conditional one, commonly used in conventional likelihood-based approaches, for analysis of LBRC-C data. Some essential properties of the joint and conditional approaches are derived and compared together. These certain properties are being scrutinized and are shown to be responsible for the superior performance of the joint approach in detecting the correct subset of covariates. Finally, Chapter 6 belongs to the analysis of the simulated example. All the results achieved in chapters 4, 5, and 6 are novel and, exclusively, contributions of the current study.

Chapter 7 is dedicated to the mapping neural networks and related questions. Particularly, we address problem P5 and establish the results we have briefly explained in the previous section.

In the end, Chapter 8 contains concluding notes and a few problems to be considered in future research.

1.7 Some Notes on Main Contributions

Before closing this introductory chapter, we would like to add some notes on the contributions of the current thesis. First, some of the problems this research settles might be regarded as a contribution to the foundations of statistical learning in the context of LBRC-C data. Specifically, the main problem of supervised learning, often, take one of the following forms: (i) density estimation, (ii) regression estimation, and (iii) classification or pattern recognition. As we explain in Chapter 3, all of these problems, as a matter of fact, are different forms of a single problem, i.e., the problem of risk minimization from empirical data. Risk minimization itself depends, directly, on the estimation of the actual probability measure, governing the distributional behaviour of the data in hand, utilizing a limited amount of sample data. In other words, supervised learning can eventually be reduced to the main problem of mathematical statistics, i.e., to construct an empirical probability measure that can consistently estimate the underlying probability measure. In Chapter 4, in addition to the non-explicit regression problem, both problems of distribution function estimation and risk minimization are effectively solved. This fundamental step paves the way for solving any supervised learning problem when data are LBRC-C.

The problem of variable selection, considered in Chapter 5, besides its theoretical novelty

and properly taking the covariate bias into account, has immediate impacts on the related practice, both for practitioners and for policymakers. Indeed, it should not be difficult to see how detecting the true risk factors of a disease, e.g., may lead to more effective use of the available financial and other resources. In terms of predictive modelling, there are situations where decisions have to be made in real time and, clearly, being able to avoid obsolete factors in producing the outcome of interest can reduce both the required time to respond as well as necessary computational expenses.

Next, we believe that fundamental research from the mathematical point of view, in the area of the artificial neural networks, deserves much more attention than it presently receives. There is no doubt that studying the applied aspects of the neural networks is also extremely important, but it should be noticed that without a profound theoretical framework there is no guarantee whether the application is actually approaching the desired destination. (Recall the metaphor mentioned earlier.) Hopefully, the theoretical scrutiny provided in Chapter 7 contributes to encouraging more research at deeper levels rather than considering individual applications per se.

Finally, both chapters 3 and 7, constitute a unique collection of results, which can provide students, researchers, and anyone with an interest in related areas with a succinct and concise introduction to statistical learning theory and the mapping neural networks without needing to explore, typically, tedious classical references. Especially, Chapter 7 gathers too many pieces of a puzzle that are scattered in various places finding which requires an incredible amount of time and effort. We know that since we learnt it the hard way. These two chapters are the by-product of our investigation of the problems of interest.

To summarize, we conclude that our contributions to knowledge, roughly, fall into three categories: (i) fundamental (theoretical), (ii) practical, and (iii) educational.

CHAPTER 2 LITERATURE REVIEW

In this chapter, we briefly look into the existing literature and summarize the results that are most pertinent to the topics studied in this thesis. The main attention will be paid to the phenomena of censoring, truncation and length bias. Due to the fairly interdisciplinary nature of our research, we thought a thematic structure for the literature review would best serve our purpose. Therefore, the chapter is divided into two independent sections and each section is devoted to a separate topic. The two sections discuss the results belonging to the following areas, respectively: (1) The learning-related literature, and (2) a general history of the length bias in different branches of science including survival analysis. Note that we did not include the mapping neural networks in this chapter since a vast portion of the discussion in Chapter 7 is already devoted to the historical aspects and to reviewing the related works that influenced the development of the neural networks.

2.1 The Learning Literature

In the present section, first, we provide a succinct historical review of the creation and advancement of statistical learning theory. Further, we will have a look at the works that are particularly related to the context of LBRC-C data.

Statistical learning theory appeared in 1960s as a purely theoretical discipline with the main objective of making a statistically sound inference about a target population, given a set of sample data. In 1990s, the invention of a new type of algorithms, called the *SVM* [Cortes and Vapnik, 1995], became a transition point for the theory of statistical learning from being a purely theoretical analysis to a practical framework capable of estimating multidimensional functions [Vapnik, 1998].

The evolution of statistical learning theory from a purely theoretical assessment of statistical principles of inference to a data analysis discipline with a rich set of practical tools was, to some extent, encouraged by the appearance of first powerful computers that were capable of conducting analysis on multidimensional, real-life data problems. Once this possibility was realized, it immediately became evident that the classical approaches towards function estimation, for the purpose of statistical inference, in low-dimensions did not reflect the problem of *singularity*, which usually occurs in the analysis of high-dimensional data. Singularity refers to situations where the variance-covariance matrix of data is ill-conditioned and, consequently, cannot be computed. This issue happens, often, when the dimension of the input

vectors becomes much larger than the sample size.

As Vapnik once phrased it, "there was *something* that could not be captured by the classical paradigm" of statistical inference [Vapnik, 1998]. The problem he referred to is what Bellman et al. [1957] called the *curse of dimensionality* in his *dynamic programming*. The desire to overcome this difficulty formed the initial inspiration for creating an alternative paradigm, which was developed by Vapnik and Chervonenkis during a course of almost three decades [Vapnik, 1998]. All the central questions and surrounding aspects of the theory were gradually established in a series of works published by Vapnik and Chervonenkis during 1960s–1990s [Vapnik and Chervonenkis, 1968, 1971, 1974a,b, Vapnik and Stepanyuk, 1978, Vapnik and Chervonenkis, 1981, 1989, Vapnik, 1992, 1995, 1998, 1999, 2006, 2019]. For this reason, statistical learning theory is sometimes referred to as Vapnik-Chervonenkis (VC) theory.

To understand the motivation behind the theory, it would be insightful to have a brief look at the earlier years and the events that preceded the birth of statistical learning. The first half of the 20th century witnessed several important events which are related to this discussion: First, the development of multiple groundbreaking algorithms for a problem that later became known as *pattern recognition*. These algorithms include Fisher's *discriminant analysis* [Fisher, 1936] and Rosenblatt's *perceptron* [Rosenblatt, 1957, 1958a,b]. Formally, the pattern recognition problem belongs to the statistical framework of estimating functions from empirical data. The reason we recall the pattern recognition problem is that, indeed, it is of primary importance in the development of learning theory, as we discuss further.

In pattern recognition, the considered class of functions consists of rather simple functions, i.e., indicator functions. Perhaps, pattern recognition is one of the simplest cases of statistical inductive inference, compared to the regression and density estimation. Nevertheless, studying this simple case became crucially important in formalizing the *generalization ability* of a general set of functions. We will see that the ability of a collection of functions to *generalize* is a closely related concept to the pivotal notion of *capacity*, in VC theory of statistical learning, and studying the pattern recognition problem served as a powerful instrument to establish the general notion of capacity for any class of functions. It was fortunate that the transition from the simple case of indicator functions to more complex ones was possible using pretty standard mathematical techniques.

Another game changer occurred in 1980s: It was suddenly noted, by Vapnik and Chervonenkis, that a generalized version of the Glivenko-Cantelli [Glivenko, 1933, Cantelli, 1933] problem would result in the same theory as the one they had developed for learning and generalization in pattern recognition. This discovery triggered motivations for promoting a comprehensive theory of learning [Vapnik, 1998].

In the sections to come, a closer look into the aforementioned set of events and their influence on the development of learning theory is provided.

Compared to inferential statistics, it has been much earlier for descriptive statistics to realize itself as a theoretically profound set of tools for studying real-world phenomena. In contrast, the history of a systematic statistical inferential infrastructure only can be traced back to the first decades of the twentieth century, when an episode of a few notable events, approximately, between 1900 and 1930, sparked off the analysis of statistical inference. Among them were the statistical hypothesis tests, significance tests, MLE, etc [Pearson, 1900, 1992, Fisher, 1920, 1922, 1925, 1992, 1970, Neyman and Pearson, 1992].

First, Fisher's unified framework of *parametric* inference for finding functional dependencies from empirical data. The problem he addressed might be generally expressed as follows: Given a set of observed data one intends to estimate the underlying statistical structure of the data distribution having some a priori information. Normally, this information is on the structure of the function to be estimated. In particular, since at the time the existence of parametric distributions had already been known, the main task was to estimate a finite number of parameters from data [Fisher, 1932].

The formulation above is rather general since it can apply to various contexts. Fisher discussed this general problem in the following settings: (i) Discriminant analysis (DA), (ii) Regression analysis, and (iii) Density estimation. The method he proposed for estimation was the *MLE*, which truly provided a powerful frame for statistical inference.

It must be noted that the MLE had been already around since a couple of hundred years before Fisher. For instance, one could find elements of it in the works of Gauss, Laplace and others, who lived long before Fisher. But Fisher was the one who popularized the use of MLE in a systematic way. However, the mathematical ground for its application was only established later. It was only in 1938 that Wilks [1938] proved a very important property of the MLE, i.e., the asymptotic behaviour of the log-likelihood ratio statistic.

The second important event was due to Glivenko [1933] and Cantelli Cantelli [1933], who worked on a similar problem but with slightly different assumptions. They proved that the empirical distribution converges to the actual distribution uniformly and does not depend on the distribution, i.e.,

$$\|\hat{F}_n - F\|_\infty = \sup_{t\in\mathbb{R}} |\hat{F}_n(t) - F(t)| \xrightarrow{a.s.} 0.$$

This is a more general setting compared to Fisher's in that no prior knowledge of distribution is required. Shortly after, Kolmogorov [1933] showed that this convergence occurs with a fast

rate, i.e., exponentially. Specifically, he showed that

$$\lim_{n\to\infty} P\{\sqrt{n}\|\hat{F}_n - F\|_\infty < \varepsilon\} = 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}\exp\{-2\varepsilon^2 k^2\}.$$

Later Smirnov [1939] also obtained similar bounds for the convergence of the empirical distribution.

Broadly speaking, there are fundamental distinctions between the core inferential principles behind Fisher's parametric approach and the one, implicitly, suggested by the Glivenko-Cantelli and Kolmogorov's theorem. They stem from two different philosophies, and this could be understood by looking at the basilar assumptions each of them makes.

Because of this difference, they initiated two different ideologies in the context of statistical inference: The well-known *parametric* or *particular* and *non-parametric* or *general* approaches. One can say that Fisher's parametric approach is narrower in its nature since the strong assumptions, based on which the theory is established, make the application domain restricted to situations where there is a considerable amount of a priori, additional information. By additional information, we mean what cannot be implied from the sample data.

What if the observed data is the only source of information? In the absence of reliable, a priori information about the actual distribution or the target function, one should naturally think of the more liberal approach, i.e., the non-parametric one, since it provides more room for applying a wider set of functions for making an inference. In other words, reducing the prerequisite assumptions opens new horizons of applicability of the inferential paradigm.

On the other hand, such machinery that is supposed to work in a very wide range of problems cannot be simple. It should, undoubtedly, be capable of dealing with a much bigger variety of occasions. But how one can determine whether a specific methodology is, actually, reliable to be employed in solving inferential problems? Statistical learning theory's answer to this question possesses two aspects:

First, any theory to be applied in the setting of the general approach must be able to provide the *necessary and sufficient* conditions for *asymptotic* optimality of the solution. Second, it should be able to provide reasonable solutions for a fixed sample size.

In other words, asymptotic behaviour solely, cannot qualify a method for real-world situations, unless it is also capable of making inference when a fixed amount of data is available. Clearly, in the context of the parametric statistics, especially with the MLE, the latter requirement sounds too much to hope for. However, Kolmogorov's discovery, concerning the convergence rate of the empirical distribution to the actual one, reinforced this hope.

Given these revelations in 1930s, indeed, the theoretical ground was already prepared for a new paradigm of inference to flourish. Nevertheless, the course of events that took place in reality, proved to have absolutely different trends. Despite the very swift development of the parametric approach, which was, basically, the only accepted paradigm until 1960s, Glivenko, Cantelli and Kolmogorov's findings did not immediately trigger thoughts about the existence of other possible theories; the ones that could have possibly been more flexible and powerful. As a matter of fact, these discoveries were mainly believed to contribute to the foundation of probability and mathematical statistics.

It took several decades, as well as a sequence of theoretical and practical events, until some researchers in the mathematical and statistical community began to deviate gradually from the mainstream settings of inference and started to question the possible existence of alternative foundations for building a new inferential system. The most marked events were related to the underlying assumptions of the parametric framework. In the following paragraphs, we briefly mention those assumptions and why they started to seem rather shaky.

In a major portion of the problems, the set of allowed functions, from which the parametric approach searches for the best dependency, usually consists of polynomials that are linear in their parameters. A classical example is the linear regression problem. There are theoretical results, such as the Stone-Weierstrass theorem, which truly justify the possibility of approximating any continuous function, with any desired accuracy, by only polynomials. This is computationally feasible in practice only if certain conditions hold. Assume that the target function is defined on the d-dimensional cube $I^d = [0,1]^d$ and is s-time differentiable. If one intends to approximate such a function by means of a polynomial with N terms, then the guaranteed accuracy is of the order $\mathcal{O}(N^{-\frac{s}{d}})$ [Vapnik, 1998]. Consequently, for approximating a non-smooth function, one has to increase the number of terms N involved in the approximation exponentially as the function's dimension grows. This is a real burden even with most of the modern ordinary computers. Recall the curse of dimensionality mentioned earlier!

Another issue is what Tukey [1960] pointed out. Real data often exhibit some peculiarities and do not obey the exact rules of the formal distributions. Hence, he suggested that one must treat data problems with specifically chosen methods capable of revealing subtleties of individual data-related situations. This fact inspired the foundations of the so-called exploratory data analysis (EDA).

Lastly, there are a few serious restrictions associated with the principal estimation machinery in the parametric approach, i.e., the MLE principle. Although the MLE enjoys some unique, desirable characteristics, it should be employed with additional care. In Chapter 5 and in

Section 8.2.1 (of Chapter 8.2), we discuss one of the possible problems of the maximum-likelihood-based inference. More precisely, the problem of sensitivity to the proper use of a priori knowledge. Next, is the lack of theoretical ground for using the MLE in finite-sample problems. While asymptotically, and in presence of some general conditions, the MLE provides optimal solutions, there is evidence suggesting its failure to achieve optimality. Finally, it has been shown that on some occasions there exists actually superior estimation strategy which uniformly provides better estimations. See, e.g., Stein's phenomenon.

Perhaps, the aforementioned circumstances were amongst the major factors behind the birth of new alternatives, including statistical learning theory. It is important to note that, besides the approach adopted by statistical learning, there were numerous other directions that resulted in other new techniques and algorithms. These attempts, though, stand at different distances from the classical ideology. In particular, some of them can be viewed as partial cures but still belonging to the mainstream ideology, while others took more radical positions and departed substantially from the accepted inferential model. Examples are robust inference [Tukey, 1960, 1962, 1977, Huber, 1964, Hampel, 1968], generalized linear models (GLM) [Nelder and Wedderburn, 1972], among others.

In Chapter 3, we delve into the foundations of statistical learning theory; in particular, we analyze the main problem of statistical inference, together with the solution statistical learning offers to overcome the limitations of the classical inferential organization. In the following paragraphs, we will focus on the studies of right-censored and length-biased data in the domain of learning theory. Note that, in what follows, we slightly extend our view to cover related works in learning theory in a broader sense. Particularly, we include some of the works in the area of artificial neural networks. Surprisingly, among different branches of learning theory, the neural networks community was the first to show interest in studying survival data. In addition, in the learning-theory-related literature, they have the largest amount of works dealing with time-to-event data. Before discussing some of these results in more detail, let us provide an overall insight on the volume of the research undertaken in various areas of learning about the different aspects of the context of our interest.

The specific type of the data we are interest in this thesis exhibit three properties: (1) Right censoring, (2) length bias, and (3) the covariate bias induced by the length bias. Beginning in 1900s, some neural network researchers started to show interest in analyzing survival data by neural networks. Since the right censoring is the most frequently seen peculiarity in survival data we realized that there have been produced a considerable amount of studies in the area of the neural networks where the right censoring is, at least, acknowledged by the authors. However, acknowledging the existence of right censoring does not mean taking it into account

during the data analysis by networks. In fact, in a good portion of these studies, the authors simply removed the censored data and advanced the analysis with the rest. Also, in some cases, the censored data were considered as missing and were *imputed* to be used for the analysis. In contrast, we were not able to find any work dealing with or even mentioning either of the length bias or the induced covariate bias.

In the statistical learning literature, the attention to survival data began to emerge more recently. Similarly, as in the neural network literature, many of the studies mention the existence of right censoring in the data analyzed. Again, not all of them provide a proper way to handle right censoring. Compared to the neural network literature, we have found, significantly, less number of works mentioning right censoring. Similarly, we could not manage to find any research on length bias data, from the statistical learning point of view. On the other hand, we have found a couple of works in *reinforcement learning* focusing on left truncated data. The left truncation is a more general form of bias and has the length bias as its special case. There has been no mention of the induced covariate bias in none of the works we have reviewed. It is also worth mentioning that none of the researches found have tackled the foundations of learning (risk minimization, density estimation, etc) in the context of either length-biased or right-censored data. That is, the predominant majority of the works are application-oriented. See, for example, *random survival forests* [Ishwaran et al., 2008] or the *SVMs* for consored data [Khan and Zubek, 2008]. Next, we provide a summary of some of the works done in the learning-related sphere.

In the neural-network literature, Snow et al. [1994] provide one of the first studies with the main interest being a certain type of time-to-event data. Their objective is to determine whether artificial neural networks would be helpful to predict biopsy results in men with abnormal prostate cancer screening test(s) and to predict treatment outcome after radical prostatectomy. However, it is not clear if the data used in this study, which were extracted randomly from a prostate-specific-antigen-based screening study database, were length biased or right censored.

Faraggi and Simon [1995] present a feed-forward neural network as the basis for a non-linear, proportional-hazards model for modelling censored survival data. According to the authors, this approach can be extended to other models used with censored survival data. The MLE is utilized to estimate the parameters of the proportional-hazards neural network. They, also, argue that these maximum-likelihood-based models can be compared, using readily available techniques such as the likelihood ratio test, the AIC, etc. The proposed model is tested on survival data of men with prostatic carcinoma. The data are right-censored, however, there is no mention of truncation or length bias.

Zupan et al. [2000] is one of the earliest studies investigating the classification problem of survival data. Roughly, the proposed approach consists of manipulating the censored data, in the first step, and then applying any existing classification method. This can be considered as an example of utilizing imputation to deal with censoring. Essentially, they estimate the probability of event for censored subjects and assign them a distribution of outcome (survival) accordingly. They, also, mention that "since most machine learning techniques do not deal with outcome distributions, the schema is implemented using weighted examples". Therefore, data manipulation is once more used to "prepare" the data for the learning process. Finally, they test their approach to build *prognostic models* for prostate cancer recurrence.

Jerez-Aragonés et al. [2003] provides a hybrid model for prognosis of breast cancer relapse based on a combination of neural networks and decision trees. They present a decision support tool that combines a novel algorithm, called *control of induction by sample division method* to select the most relevant prognostic factors (variable selection). Then, a system composed of different neural networks topologies takes the selected variables as input to produce the final survival class. Unfortunately, it is not clear how the censored data are handled.

Bøvelstad et al. [2007] compares some of the most frequently used regression approaches in statistical learning to the regression analysis of high-dimensional survival data. The main motivation is predicting the survival time of patients with cancer based on gene expression profiles. Apart from the difficulties related to high-dimensional genome-wide expression data, they mention censoring as another source of challenge. All the methods considered are combinations of some dimensionality reduction method such as the principle component analysis and the Cox model. Hence, they do not offer any special treatment of censoring from the learning point of view. Eventually, they discuss neither the length bias nor the truncation.

Next, Witten and Tibshirani [2010] provides a comparative study of statistical learning methods applied to high-dimensional survival data. While censoring is discussed by the authors, no truncation or length bias is mentioned there. Note that the authors provide a quite thorough and useful list of references in the statistical learning domain with interest in survival data.

Steele et al. [2018] studies prognostic modelling based on electronic health records (EHR) data and compares "conventional epidemiological approaches" to machine learning ones such as the random forests. They simply remove the data related to the censored subjects. Length bias or truncation is not of their interest.

Macías-García et al. [2020] is another research that studies time-to-event data in connection

with breast cancer recurrence. The data are subject to type I censoring, however, the time-to-event or survival time for the censored subjects at the end of the study are replaced by their censoring time.

Similar to Bøvelstad et al. [2007], Spooner et al. [2020] studies and compares the performance of some machine learning models on high-dimensional survival data with non-informative censoring. The models considered are either of the following forms: (i) A combination of some variable selection or dimensionality reduction methods with the Cox proportional-hazards (CPH) model, such as a penalised Cox Regression model with either the Least Absolute Shrinkage and Selection Operator (LASSO), the Elastic Net, or the Ridge penalties. (ii) Ensemble methods with or without the CPH model, including *boosted* CPH models, such as the Cox model with likelihood-based boosting (CoxBoost), Cox model with gradient boosting (GLMBoost) and Extreme Gradient Boosting (XGBoost) with tree-based and linear model-based boosting, and the random forests. To deal with censoring, Spooner et al. [2020] employ the Cox model. No truncation, including length bias, is considered. It is worth mentioning that a valuable list of references related to statistical learning of the time-to-event is provide by the authors.

The number of the works in statistical or other branches of learning theory that somehow are related to survival data is huge. Fortunately, we were able to look into numerous researches, in learning theory, for the purpose of this literature review. However, due to the number of the works considered, it is impossible for us to mention all of them in this section. Nonetheless, we are able to summarize them as follows. While the aforementioned list of the works is just a tiny, and non-exhaustive list of examples, where the interest rests in time-to-event data, it reflects the major tendency in the related areas of research. Just as we have seen in the examples above, the majority of the researches on survival data, from the learning point of view, is restricted to considering censoring, solely, with no mention of the truncation or length bias at all. We were able to find few articles with interest in left-truncated (but not length-biased) data, in the domain of reinforcement learning. Particularly, Daskalakis et al. [2019], and Daskalakis et al. [2020] study the regression problem when data are left-truncated. In addition to the motivation of these works which are different from the one of ours, there is, also, no censoring involved in these researches.

To conclude, we add a few more examples to the already mentioned works. Note that these are all connected to the analysis of survival data from the learning point of view, where the length bias or truncation were not of interest but censoring, at least, has been mentioned: Graf et al. [1999], Heagerty et al. [2000], Nguyen and Rocke [2002], Hothorn et al. [2004], Li and Gui [2004], Li and Li [2004], Shivaswamy et al. [2007], Schumacher et al. [2007], Bittern

et al. [2007], Ishwaran et al. [2008], Khan and Zubek [2008], Goldberg and Kosorok [2012], Binder [2013], Padhukasahasram et al. [2015], Ranganath et al. [2016], Martinsson [2016], Vock et al. [2016], Lee and Lim [2019], Nemati et al. [2020], De Laurentiis and Ravdin [1994], Liestøl et al. [1994], Biganzoli et al. [1998], Ripley and Ripley [2001], Joshi and Reeves [2006], Chi et al. [2007], Montes-Torres et al. [2016], Nilsaz-Dezfouli et al. [2017], Ching et al. [2018], Kvamme et al. [2019], Bengio [2019], Tilman [2020].

2.2 A History of Length Bias

In the present section, we focus on the literature related to the length bias. It is important to note that the occurrence of the length bias and right censoring is not restricted to a specific field. The following examples belong to various disciplines and, consequently, some of them included a considerable amount of domain-related technical details, which were not completely relevant to the topic of our interest. For this reason, such details were mostly skipped.

Another thing worth mentioning is that because length bias appears in situations belonging to different areas of research, it was rarely discussed as the central subject of interest and in a unified and systematic way. Hughes and Savoca [1999] argued that this lack of attention to this issue may be explained by the fact that the earliest and most frequent applications of duration analysis, in general, appeared in experimental designs where one has essentially more control over inclusion of relevant cases to the study. Finally, we would like to mention that the core intention of writing this section, perhaps, might be summarized as to emphasize the importance of the problem of length bias in data analysis by showing that this is a ubiquitous problem that has been revisited frequently, in plenty of situations, and in diverse areas of research.

Studying the phenomenon of length bias dates back, at least, to the first decades of the 20th century. The rise of length-biased data in practice can be probably attributed to biology and, especially, the works of Wicksell on the famous corpuscles problem. Wicksell [1925, 1926] investigated the bodies of corpuscles formed in tissues of different human and other animals' organs. It was known at the time that the corpuscle bodies varied, considerably, in terms of both size and numbers, even within one single organ. Accordingly, Wicksell's main objective was to detect the distribution of the size and numbers of the corpuscle bodies in individual organs, as well as, their distribution in distinct organs and individuals. Back then, and in absence of today's modern technology, to do the measurements, Wicksell had no option but cutting the tissue and then quantifying the corpuscles' sizes and counts on the two-dimensional, cross-sectional surfaces of the cut tissue. This sampling method is

length biased as corpuscles with larger diameters are more likely to be cut and observed in the sample. Moreover, this "two-dimensional sampling" of the de facto three-dimensional corpuscles introduced another kind of bias as well: The diameter of an arbitrary sampled corpuscle was less than or equal to its true diameter if the diameter was defined as the longest distance between any two points of the corpuscle. A year later, Wicksell [1926] studied a similar problem and succeeded to extend the previously obtained results to the case of ellipsoid corpuscles.

Fisher [1934] and Neyman [1955] also considered the problem of estimation of the incidence frequency utilizing length-biased data collected, from the prevalent population.

McFadden [1962] investigated the length of intervals between a sequence of events generated according to a *stationary point process*. Let $\left(X_t(\omega), t \in T\right)$ be a stationary stochastic process with T being a subset of the naturals $\mathbb{N}$. Define the sequence of the intervals between the events as $I_t := X_{t+1} - X_t$. Now, if one samples intervals at a random moment r, then, apparently, the length of the chosen intervals is subject to length bias. McFadden was interested in differences between the original set of intervals $\left(I_t, t \in T\right)$ and the sampled ones.

Blumenthal [1967] was interested in estimating the mean life of electron tubes. An electron tube or valve is a device that controls electric current flow between electrodes, in a high vacuum, to which an electric potential difference has been applied. The particular sample he had access to consisted of the tubes that had already been working for a certain amount of time. Similar to the sampling design employed in the cross-sectional, prevalent-cohort studies, Blumenthal's sample was biased since longer-lasting electron tubes enjoyed more chances of being selected in a sample. Blumenthal considered two types of systems. First, systems with "finite" age, and second, those with extremely long lifetimes, called *equilibrium*. Since the backward and forward recurrence times were assumed to be distributed identically, the question was whether the knowledge of the backward recurrence time suffices to estimate the mean lifetime of a tube. From the practical point of view, this is a sensible question as the backward recurrence time can be measured immediately after a tube is sampled, whereas measuring the forward recurrence time requires waiting until the failure occurs. Blumenthal's answer though showed a higher efficiency for the estimation process with both backward and forward recurrence times taken into account. He explained that the estimates will have less variance due to the fact that backward and forward recurrence times are not independent. Special attention has been paid to the gamma and Weibull families of distributions.

Cox [1969] discussed, among others, the process of assessing the quality of a fabric through the length of its fibres. It was supposed that the fabric's quality was positively correlated

with the length of its fibres. Therefore, to control the quality, fibres were randomly sampled by a comb and measured subsequently. This sampling mechanism favoured longer fibres by giving them higher chances of being selected. The moments of the distribution of the length-biased sampled fibres and their relation with those of the unbiased distribution were targeted in the study. In addition, it was shown that the length-biased sampling provided more efficient estimates of the upper tail of the distributions when the unbiased distribution of the fibre's length was assumed to be either Log-Normal or Gamma. This is not surprising because the upper tail of the distribution is, in fact, where the length-biased sampling avidly tends to choose from. It is worth mentioning that since the applications motivating the aforementioned studies did not involve censoring, no correction for considered data has been considered.

After the basic theory was established, the number of applications raised dramatically over time. For example, Goldsmith [1967] realized that there were equivalents to the original Wicksell's anatomy problem in physics: The true particle size obtained from a thin section of an object is distorted and, as a result, the distribution of the particle sizes cannot be immediately observed. Goldsmith provided calculations for finding true distribution in the aforementioned case. Similar issues were encountered in astrophysics related to the problem of cataloging galaxies. Likewise in biology, Smith et al. [1969] dealt with an analogous problem in a study of carcinogenesis. Carcinogenesis, also called oncogenesis or tumorigenesis, refers to the process of the transformation of normal cells into cancerous ones.

Zelen's works in the area of screening tests and randomized trials, in late 60s and early 70s, dealt with general sampling biases. These works, some of which were coauthored with other researchers, played a considerable role in raising awareness and recognition of the biased sampling's implications in medical and epidemiological studies. For instance, one may refer to Feigl and Zelen [1965], Zelen [1969, 1973, 1979], for some related results. Particularly, in a later work, Davidov and Zelen [2001] pointed the effect of size-biased sampling in the investigation of the relative risk of diseases based on family history. According to the authors, the familial risk of having a disease is, usually, assessed through case-control studies based on databases consisting of a collection of family histories of cases typically assembled as a result of one family member being diagnosed with the target disease. They argued that sampling based on family registries is size biased because larger families are found in registers with higher probabilities (i.e., proportional to their size).

Lagakos et al. [1988] carried out an important research on AIDS and related issues, where the data were sampled through a cross-sectional prevalent-cohort design. In this case, the original data were right truncated but using a reverse time transformation they were converted into

left-truncated data, so that the techniques available for left truncation could be applied to analyze the data.

The length-biased methodology has also been applied to data analysis in multiple studies in the domain of Alzheimer's and dementia. For example, Gao and Hui [2000] used a two-phased sampling scheme to estimate incidence of dementia. On the other hand, in an analysis related to Alzheimer's, Stern et al. [1997] used data which were extracted from prevalent cohort, but no correction for length bias was accounted. Wolfson et al. [2001] showed that, when length bias is taken into account properly, the median survival lifetime of individuals with dementia is considerably shorter than it was previously estimated. This study was based on the data from the Canadian Study of Health and Aging (CSHA).

As already seen, aetiology is an area where length-biased data are quite prevalent. Interestingly, the setting in which length bias is discussed in aetiology differs from the setting adopted in conventional lifetime analysis in that the target variable in aetiology is the *prevalence (proportion)* of a characteristic or an attribute among individuals who already have developed a certain disease. This is, unquestionably, an important issue in immunogenetics since it helps establish whether a characteristic is related to the aetiology of the disease, i.e., if it plays a causal role, or it is of prognostic importance, in which case the characteristic could be implied from the disease and not vice versa. An example of an aetiologic study with length-biased data is Simon [1980].

As well as in biomedical sciences, the length bias is of great interest in the theory of *renewal processes*. A renewal process can be described as follows: Assume that a number of objects from a particular population are put in operation at the present moment. Whenever an object fails, a new object from the same population will replace it. Now, suppose that, at some time r in distant future, an inspector collects the data of the operating-at-that-moment objects and monitors the objects for a fixed amount of time s after collection. The collected data contain information about the age of each object at the moment of collection. The variable of interest is the mean lifetime of the population while available sample is limited to the objects in use at the collection time. Assuming that r is large enough such data are length biased. Moreover, if $s < \infty$, then the sample is right censored as well. Renewal processes have been investigated in Winter and Földes [1988], among others.

Hughes and Savoca [1999] studied the impact of legal reforms on the duration of legal disputes over medical malpractice. The sample used was obtained from insurance claims during a period of 4 years ($1985 - 1989$) in the USA. The data were both length biased and right censored and the authors proposed corrections for both issues. Interestingly, the result of the study revealed that the "English rule", i.e., a rule that requires the loser at trial to cover

all legal expenses, is the only factor among several competing factors which shortens the length of a dispute. Moreover, it was illustrated that failing to account for length bias might result in an absolutely inverse conclusion, i.e., the English rule may seem to lengthen the time needed for settlement and litigation.

Often, length bias occurs when the variable of interest involves time or sampling mechanism might be affected by time. In economics, for instance, while time might not be the main objective of investigation, it can play a role in sampling. The influence of length-biased sampling on contingent value studies, which are used to quantify the value of non-monetary variables such as environmental commodities and non-traded goods, was explored by Nowell et al. [1988]. Sampling users of such commodities (e.g., a fishing resort) requires being on site while individuals are in the middle of an activity. This implies that those users who tend to spend more time, and thus put more value on the contingents, are more likely to be sampled. Nowell and Stanley [1991] noted an akin length bias in mall intercept surveys. For example, they pointed out that selecting shoppers inside a shopping center rather than sampling at the entrance exhibits some properties of length bias. Similarly, there could be differences in tendencies of being sampled between individuals who visit many different stores in comparison to those who spend a longer time in a single shop. As a final example of applying length bias methodology in economics one may think of studying the length of spells of unemployment. A related reference can be found in an unpublished script by De Uña Álvarez [2001].

To conclude this chapter, we would like to emphasize that while the provided list of the works considered in this section is not comprehensive, it is sufficient to see the ubiquity of the length bias, and consequently, its importance as a research context in the field of data analysis. Nonetheless, as we have seen in the previous section of the present chapter, the issue has not received much attention, especially, from the learning theory community. We hope that the next chapters shed some light on the problem, in particular, from the statistical learning viewpoint.

CHAPTER 3 PRELIMINARY KNOWLEDGE

In this chapter, we provide the background knowledge necessary for the rest of the thesis. Particularly, as the main object of interest is statistical learning from a certain type of incomplete data, which is often encountered in the analysis of time-to-event or survival data, here we review the following two topics:

1. Foundations of statistical learning theory, and
2. Time-to-event or survival analysis.

Sections 3.1 and 3.3 are dedicated to these topics, respectively.

3.1 Foundations of Statistical Learning Theory

Statistical learning theory, as one of the maturest branches of data science and machine learning, has evolved around the main problem of statistics, i.e., making sound statistical inference about a target population based on a limited amount of sample data. While adopting some fundamental concepts from statistics, statistical learning theory exploits some techniques from functional analysis in order to provide a coherent mathematical framework for solving data-related problems in real-world situations. It has, also, extended the power of classical tools of statistics to new areas of application by successfully consolidating statistics and computer-related sciences. This section explores the foundations of statistical learning theory and presents a condensed overview of its main results and their connection with some previously achieved results in mathematics, such as the well-known Glivenko-Cantelli theorem and the law of large numbers.

3.1.1 Inference in Statistical Learning Theory

First of all, let us introduce the main problem of statistical learning theory, according to VC theory, which is the primary theoretical framework of this research. In what follows, we formulate the main problem of *supervised* learning since the weakly-supervised learning context considered in this work is, as a matter of fact, a sub-category of supervised learning.

Assume that there exists a *generating mechanism* that *randomly* generates *independent* input vectors $\mathbf{X} \in \mathbb{R}^d$ according to an unknown, fixed probability $P_{\mathbf{X}}(\mathbf{x})$. More precisely, $\mathbf{X}$ is a measurable function $\mathbf{X} : \mathcal{X} \to \mathbb{R}^d$, where $\mathcal{X}$ is a sample space of a probability space $(\mathcal{X}, \Sigma_{\mathcal{X}}, P_{\mathbf{X}})$, and $(\mathbb{R}^d, \mathcal{B}_d)$ is the Borel measurable space.

Also, there is a "*supervisor*" whose job is to *assign* an output $Y \in \mathbb{R}$ to each generated input $\mathbf{X}$, through an unknown, fixed, stochastic procedure $P_{Y|\mathbf{X}}$.

Additionally, there exists a *machine* whose objective is to realize how to, properly, assign an output to each input, as does the supervisor, by observing a *limited amount of data*, called the *training data* or *sample set* $\mathcal{D}$. Therefore, the training set is a finite collection of realized input vectors, together with their corresponding outputs assigned by the supervisor, which where collected *independently*. That is,

$$\mathcal{D} = \left\{ (\mathbf{x}_i, y_i) \in \mathbb{R}^{d+1} \mid i = 1, \ldots, n \right\}. \tag{3.1}$$

Note that a "proper" output is one that is *close*, in consonance with some similarity measure, to the output the supervisor would appoint to $\mathbf{x}$. However, the machine is not aware of the supervisor's assignment mechanism and the only information available for the machine to come upon the outputs is the sample data $\mathcal{D}$.

It is important to note that the provided samples come from the joint distribution $P_{\mathbf{X},Y}$, rather than the conditional $P_{Y|\mathbf{X}}$. (See Figure 3.1.)

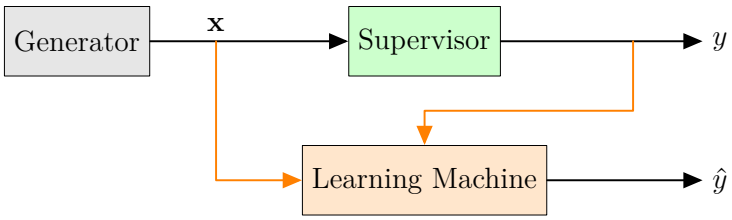

Figure 3.1 **Vapnik's Model of Learning from Examples.** Vapnik's model consists of three elements: a *generator*, a *target operator* or *supervisor*, and a *learning machine*. During the training procedure, the learning machine is provided by a finite number of $(\mathbf{x}, y)$, labeled by the supervisor, based on which the machine produces the estimate $\hat{y}$. (Concept borrowed from Vapnik [1998].)

Formally, the underlying relation between $\mathbf{X}$ and Y is called a *(functional) dependence* or a *hypothesis*. In practice, the *learning machine* is at liberty to choose a dependence, say $h_{\boldsymbol{\theta}}$, only from a *predetermined hypothesis space* $\mathcal{H}_{\boldsymbol{\Theta}}$, which is a *distinguishable* family of parametric hypotheses. More precisely, $\mathcal{H}_{\boldsymbol{\Theta}} := \{h_{\boldsymbol{\theta}} : \mathbb{R}^d \to \mathbb{R} \mid \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, where $\boldsymbol{\Theta}$ is an arbitrary index set.

Generally, $\mathcal{H}_{\boldsymbol{\Theta}}$ is assumed to be a vector space, usually, equipped with some additional structure, such as a norm or an inner product, in order to facilitate, for example, defining a (dis)similarity measure over the vectors. A common choice is a Hilbert space.

Since we assume a one-to-one correspondence between the hypothesis space $\mathcal{H}_{\boldsymbol{\Theta}}$ and its index

set $\boldsymbol{\Theta}$, with a slight abuse of notation, we can talk of $\boldsymbol{\theta}$ in place of $h_{\boldsymbol{\theta}}$. Accordingly, $\boldsymbol{\Theta}$ will be called the hypothesis space.

The scenario described above, gives rise to two completely different inferential models:

1. The first one involves a two-step inferential process, namely, an *inductive* step, which is to identify the supervisor's general "wisdom" from the particular set of training data, and subsequently, a *deductive* step consisting of applying the revealed wisdom to arbitrary future cases.

2. The second model requires a type of inference that involves no generalization. More precisely, the learning machine is not interested in discovering the general wisdom of the supervisor; in contrast, its goal is to, merely, infer the outputs for the members of a certain subset $S \subset \mathbb{R}$. This type of inference is called *transductive* inference.

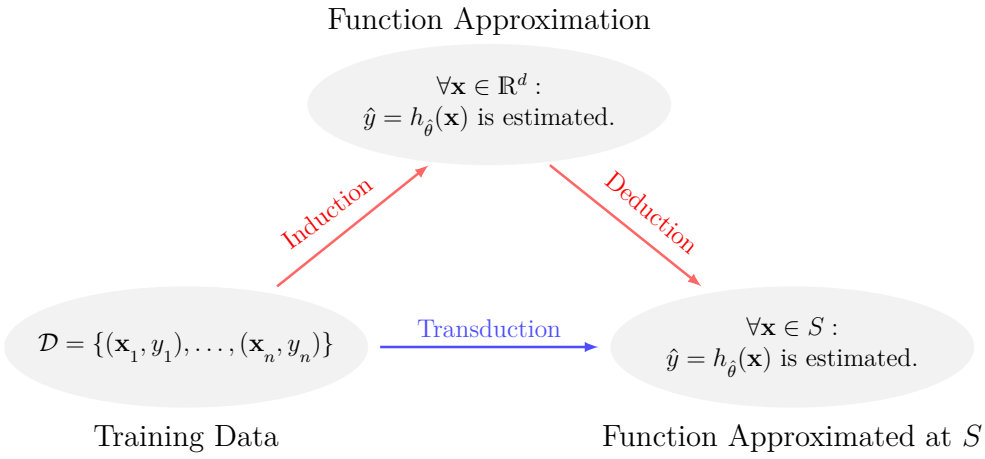

Figure 3.2 **Induction vs. Transduction.** Here, the two potential types of inference, as Vapnik described, are depicted: One uses the training data, $\mathcal{D}$, to estimate the functional dependence $h_{\boldsymbol{\theta}}$ at the entire input space, through an *inductive* inferential scheme and, then, applies the estimated function $h_{\hat{\boldsymbol{\theta}}}$ to *deduce* the values of function at the points of interest, i.e., points belonging to S, for example. In contrast, the *transductive* approach uses the training data to, directly, infer the values of $\hat{y}$ at each point of S. (Concept borrowed from Vapnik [1995].)

While the former problem is considered to be harder compared to the latter, it turns out that the general principles, by means of which these classes of problems might be solved are alike. Clearly, the reason behind the difficulty of solving problems of the first type is that it necessitates estimating the function $h_{\boldsymbol{\theta}}$ over the entire input space before being able to employ it for further *predictions*. In the current study, we consider the first approach, which involves a two-step inference, i.e., induction and deduction.

3.1.2 Inductive Principle of Estimation by Risk Minimization

Recall that the main goal of the learning machine is to find a hypothesis $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ that possesses the optimal performance in assigning an output Y to any arbitrary input $\mathbf{X}$, compared to other admissible hypothesis. Particularly, it is desirable for the chosen hypothesis to be able to *generalize*, i.e., to be able to label unseen examples. This is usually achieved by minimizing the *expected amount of mistakes* any hypothesis $\boldsymbol{\theta}$, belonging to the set of admissible functions, does if it was to be employed for assigning the outputs. The meaning of "mistake" is highly context-dependent. To, systematically, assign an error to each output, generated by a hypothesis $\boldsymbol{\theta}$, one uses a loss function. Hence, we shall first formalize this concept:

Definition 1 (Loss Function)**.** *Let $\boldsymbol{\Theta}$ be the hypothesis space and $\mathbf{X}, Y$ denote the input and output, respectively. Then, a loss function L, defined for any hypothesis $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and a realized vector $\mathbf{z} = (\mathbf{x}, y)$, is a measure of dissimilarity between $h_{\boldsymbol{\theta}}(\mathbf{x})$ and y, i.e.,*

$$\begin{aligned} L : \boldsymbol{\Theta} \times \mathbb{R}^{d+1} &\to \mathbb{R}_{\geq 0} \\ (\boldsymbol{\theta}, \mathbf{z}) &\mapsto L_{\boldsymbol{\theta}}(\mathbf{z}), \end{aligned}$$

where $L_{\boldsymbol{\theta}}$ is $P_{\mathbf{Z}}$-integrable, i.e., $\int L_{\boldsymbol{\theta}}(\mathbf{z}) \, \mathrm{d}P_{\mathbf{Z}}(\mathbf{z}) < \infty$.

Given a loss function, we are able to define the next important concept that generalizes, in the explained sense, the amount of loss associated with a hypothesis $\boldsymbol{\theta}$. This generalized loss is called *(expected) risk* and is what the learning machine "desires" to minimize, over the set of admissible hypothesis $\boldsymbol{\Theta}$, in order to select the optimal hypothesis.

Definition 2 (Expected Risk Functional)**.** *Suppose $L_{\boldsymbol{\theta}}$ denotes a particular loss function. Then, the expected risk is a functional $R : \boldsymbol{\Theta} \to \mathbb{R}$ that assigns, to each hypothesis $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, a certain value $R(\boldsymbol{\theta})$, called the expected risk associated with $\boldsymbol{\theta}$, which is defined by*

$$\begin{aligned} R(\boldsymbol{\theta}) &:= \mathbb{E}_{\mathbf{X},Y} \left\{ L_{\boldsymbol{\theta}}(\mathbf{X}, Y) \right\} \\ &= \int L_{\boldsymbol{\theta}}(\mathbf{u}, v) \, \mathrm{d}P_{\mathbf{X},Y}(\mathbf{u}, v). \end{aligned} \tag{3.2}$$

In general, the risk functional R depends on the distribution $P_{\mathbf{X},Y}$, as well as, the loss function $L_{\boldsymbol{\theta}}$. Nevertheless, since both of them are assumed to be fixed during the learning process we can consider the expected risk as a function(al) of only the hypothesis, thus the single argument $\boldsymbol{\theta}$ for the functional R.

Now, having the definition of the expect risk, one is able to formalize the main problem of statistical learning theory as follows.

The Learning Problem. *Let $P_{\mathbf{X},Y}$ be the fixed, unknown, joint probability measure underlying the observations and $\boldsymbol{\Theta}$ be the hypothesis space. Then,* learning *is defined as solving the minimization problem*

$$\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \Big\{R(\boldsymbol{\theta})\Big\},$$

where an i.i.d. sample $\mathcal{D}$, defined by equation (3.1), is given.

Although, minimizing the expected risk functional over the hypothesis space seems a reasonable tool to be used for solving the problem of finding the best hypothesis in $\boldsymbol{\Theta}$, there is no actual way to compute it because of its dependence on the *unknown* distribution $P_{\mathbf{X},Y}$ (see equation (3.2)). Alternatively, one may resort to a "suitable" sample-based estimator of $R(\boldsymbol{\theta})$. One natural possibility is given in the next definition:

Definition 3 (Empirical Risk Functional)**.** *Let $\mathcal{D}$ be the sample defined by equation (3.1), and $L_{\boldsymbol{\theta}}$ be a loss function. Then, the functional $\hat{R}_n : \boldsymbol{\Theta} \to \mathbb{R}$ is called the empirical risk and is defined as*

$$\hat{R}_n(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n} L_{\boldsymbol{\theta}}(\mathbf{x}_i, y_i).$$

Definition 3, enables one to solve the learning problem based on empirical data. That is, by minimizing the empirical risk $\hat{R}_n(\boldsymbol{\theta})$, instead of $R(\boldsymbol{\theta})$. As mentioned earlier, while the empirical risk functional seems to be an intuitive and natural surrogate for the expected risk, the adequacy of such a substitution has to be investigated.

In fact, the theoretical justification of this replacement comprises a considerable amount of the theory developed in statistical learning theory. Similarly as in classical statistics, the goodness of the empirical risk, in order to be applied for solving the learning problem in place of the expected risk, is studied by examining its *consistency*. This makes sense as the empirical risk functional is a data-based estimate of the expected risk functional.

Before discussing the consistency, there is one more point to be noted: The ERM principle had been being applied in statistics long before the emergence of learning theory. For example, it should not be difficult to see that the least-squares regression or the maximum likelihood estimation methods are both special cases of the empirical risk minimization; the distinction barely lies in the choice of the loss function.

3.1.3 Consistency of the Empirical Risk Minimization

In statistics, asymptotic theory is the standard framework for assessing the general adequacy of a sample-based statistic. *Asymptotic consistency* is usually a desirable property for an estimator.

There is an important point about the notion of consistency that must be clarified. Oftentimes, the usefulness of asymptotic properties is being criticized for being "unrealistic" since it considers the behaviour of an estimator in the idealistic occasion where an *unlimited* amount of data is available. Perhaps, this confusion stems from a subtle misinterpretation of what, actually, consistency implies. Put it non-technically, asymptotic consistency is merely a *bare minimum* for an estimator to be regarded as reliable. It is important to note that considering the asymptotic consistency as a *sufficient* condition for the accuracy of an estimator is, simply and unduly, confusing its actual meaning. Indeed, consistency should be viewed as a *necessary* condition rather than a *sufficient* one.

Before further discussing the consistency, let us remind you that in the upcoming chapter we use two key concepts from probability theory and statistics, namely, *convergence in probability* and *almost sure convergence* of sequences of random variables. One may easily find the definitions in almost any standard text in probability and statistics or related areas. Her are a few examples: Casella and Berger [2002], Wilks [1943] or Vapnik [1998]. Here, convergence in probability and almost surely will be denoted by $\xrightarrow{p}$ and $\xrightarrow{a.s.}$, respectively.

In the following paragraphs, first, we recall the classical definition of a consistent estimator. After that, a closely related but slightly different notion will be discussed. This new concept is called *nontrivial consistency* and is the one used in statistical learning theory.

Definition 4 (Consistency of an Estimator)**.** *Let* $\left\{P_{\gamma} : \gamma \in \mathbf{\Gamma}\right\}$ *be a family of parametric probability measures defined on a measurable space* (Ω, Σ) *and* X_i*,* $i = 1, 2, \ldots, n$*, be i.i.d. random variables distributed according to* P_{γ}*. Then,* $\hat{\gamma}_n := \hat{\gamma}_n(X_1, X_2, \ldots, X_n)$ *is called a consistent estimator of the parameters* γ *if and only if* $\hat{\gamma}_n$ *converges to the actual value of* γ*, in probability, i.e.,*

$$\hat{\gamma}_n \xrightarrow{p} \gamma, \text{ as } n \to \infty, \quad \forall \gamma \in \mathbf{\Gamma}.$$

(For the definition of consistency see Ibragimov and Has'minskii [1981], among others.)

For example, the so-called *sample mean* and *sample median* are consistent estimators of the population mean and median, respectively; provided that they exist and are well-defined.

Further, we define a particular type of consistency, which is of interest in the context of learning theory and is slightly stronger than the classical consistency. The new concept is

called *nontrivial consistency*.

Definition 5 (Nontrivial Consistency of ERM)**.** *Given a hypothesis space* $\boldsymbol{\Theta}$*, define the subset* $\boldsymbol{\Theta}_c \subseteq \boldsymbol{\Theta}$ *as*

$$\boldsymbol{\Theta}_c := \{\boldsymbol{\theta} \in \boldsymbol{\Theta} \mid R(\boldsymbol{\theta}) > c\}, \quad c \in \mathbb{R}.$$

Then, the ERM is said to be nontrivially (or strictly) consistent over $\boldsymbol{\Theta}$ *if for any real* c *holds*

$$\inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_c} \left\{\hat{R}_n(\boldsymbol{\theta})\right\} \xrightarrow{p} \inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_c} \left\{R(\boldsymbol{\theta})\right\}, \qquad \text{as } n \to \infty.$$

The reason for calling it "nontrivial" might be explained as follows: Suppose that $\boldsymbol{\Theta}$ be a hypotheses space, over which the ERM is not consistent. Now, extend $\boldsymbol{\Theta}$ by adding a function $\boldsymbol{\theta}_0$, for which we have that

$$L_{\boldsymbol{\theta}_0}(\mathbf{x}, y) < \inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\{L_{\boldsymbol{\theta}}(\mathbf{x}, y)\right\}, \quad \forall (\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}.$$

Clearly, this makes the ERM consistent over the extended hypothesis space $\boldsymbol{\Theta} \cup \{\boldsymbol{\theta}_0\}$ as both the empirical and expected risk reach their minimum at $\boldsymbol{\theta}_0$, independently from the data distribution and sample size n. The nontrivial consistency, in Definition 5, excludes such "trivial" cases.

In the sequel, we discuss the conditions under which the ERM principle constitute a nontrivially consistent estimator for the expected risk. First of all, one of the key achievements in the theory of learning will be introduced. This is a theorem proved by Vapnik and Chervonenkis [1989] that establishes the *sufficient and necessary* conditions for nontrivial consistency of the ERM.

The nontrivial consistency of the ERM is established utilizing the following two related stochastic processes. First, consider the *one-sided* empirical process defined by the following random variable:

$$\xi_n^+ := \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\{R(\boldsymbol{\theta}) - \hat{R}_n(\boldsymbol{\theta})\right\}_+,$$

Where

$$u_+ := \begin{cases} u & \text{if } u > 0, \\ 0 & \text{otherwise.} \end{cases}$$

In fact, ξ_n^+ and some of its particular properties were crucial to investigation of the nontrivial consistency of the ERM. Specifically, they proved that if ξ_n^+ converges, in probability, to zero

as n grows to infinity, then the ERM is nontrivially consistent over the hypothesis space $\boldsymbol{\Theta}$, and conversely. In other words, convergence of the one-sided empirical process ξ_n^+ provides a necessary and sufficient condition for the nontrivial consistency of the ERM. Further, we will introduce this theorem in more details.

Theorem 1 (One-Sided Uniform Law of Large Numbers)**.** *Let* $r_1, r_2 \in \mathbb{R}$ *and* $\boldsymbol{\Theta}$ *be a set of hypotheses such that for any* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ *holds* $r_1 \leq R(\boldsymbol{\theta}) \leq r_2$*. Then, the ERM is nontrivially consistent over* $\boldsymbol{\Theta}$*, iff*

$$\lim_{n\to\infty} \mathbb{P}\left\{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left[R(\boldsymbol{\theta}) - \hat{R}_n(\boldsymbol{\theta})\right] > \varepsilon\right\} = 0, \quad \forall \varepsilon > 0. \tag{3.3}$$

Therefore, according to theorem 1, verifying the nontrivial consistency of the ERM might be replaced by evaluating the validity of the limit (3.3). Notice that (3.3) actually gives the convergence in probability of ξ_n^+, because

$$\mathbb{P}\left\{|\xi_n^+| > \varepsilon\right\} = \mathbb{P}\left\{\left|\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left[R(\boldsymbol{\theta}) - \hat{R}_n(\boldsymbol{\theta})\right]_+\right| > \varepsilon\right\} = \mathbb{P}\left\{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left[R(\boldsymbol{\theta}) - \hat{R}_n(\boldsymbol{\theta})\right] > \varepsilon\right\}.$$

It was mentioned before that although consistency of any estimation procedure does not necessarily imply its quality, it should be a desirable property for any estimator to possess. As a result of the equivalence established in the theorem above, equation (3.3) may obviously play a crucial role in learning theory and is worth a closer look. To this end, we will introduce two stochastic processes, through which the validity of the convergence (3.3) will be verified later.

Now, consider the following *two-sided* empirical process:

$$\xi_n := \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\{\left|R(\boldsymbol{\theta}) - \hat{R}_n(\boldsymbol{\theta})\right|\right\}, \quad n = 1, 2, \ldots .$$

If $|\boldsymbol{\Theta}| < \infty$, i.e., the hypothesis space contains only a finite number of functions, then ξ_n converges in probability to zero as n increases. To see this, it is enough to notice that $\hat{R}_n(\boldsymbol{\theta}) \xrightarrow{p} R(\boldsymbol{\theta})$, according to the law of large numbers (LLN), which in turn implies $\xi_n \xrightarrow{p} 0$, as $n \to \infty$. This can be interpreted as the $|\boldsymbol{\Theta}|$-*dimensional (uniform) law of large numbers*, which describes the LLN in a vector space of a finite dimension. In this case, it requires simultaneous convergence in probability in every coordinate. In the current case, each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ corresponds to a coordinate of the vector space [Vapnik, 1995].

The problem arises when $|\boldsymbol{\Theta}|$ is infinite. In fact, in the infinite occasion, ξ_n does not necessar-

ily converges to zero, in probability, as the sample size tends to infinity. Hence, it turns out that the hypothesis space's properties impact the potential generalizability of the uniform law of large numbers to functional spaces, broadly speaking. Accordingly, the next step is to formalize the hypothesis space's properties contributing to the uniform LLN to hold.

3.1.4 Capacity or Complexity of the Hypothesis Space

A hypothesis space must contain *diverse* enough hypotheses $\boldsymbol{\theta}$'s in order to make the model a good fit for the training data. In other words, the richer the space $\boldsymbol{\Theta}$, the higher likelihood that the learning machine contains a dependence that is a good fit for the training data. Therefore, the measure of diversity plays a crucial role. Nonetheless, the choice of the complexity measure is not necessarily obvious. This section, is devoted to this topic from the statistical learning point of view.

VC theory employs a type of *entropy* in order to develop the notion of complexity for a set of admissible hypothesis. Note that, multiple concepts of entropy had been introduced, and in different contexts, before the VC entropy. Despite being defined and applied in seemingly unrelated ways across different fields, all of these definitions can be, to a very good extent, boiled down to a unifying core idea underlying all of them. For example, one may refer to Vapnik [1995] for the difference between the VC entropy and the well-known *metric entropy*.

Before introducing the VC entropy, let us simplify our notation, for the remainder of this chapter:

- Denote $\mathbf{Z} := (\mathbf{X}, Y)$. Subsequently, the training data, defined by equation (3.1), is of the form $\mathcal{D} = \{\mathbf{z}_i \in \mathbb{R}^{d+1} \mid i = 1, \ldots, n\}$. Note that everywhere throughout this thesis n denotes the sample (or training) size, unless otherwise stated.
- Given $\mathcal{D}$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, define $q(\boldsymbol{\theta}, \mathcal{D})$ as the vector of the corresponding losses, i.e.,

$$q(\boldsymbol{\theta}, \mathcal{D}) := \Big(L_{\boldsymbol{\theta}}(\mathbf{z}_1), L_{\boldsymbol{\theta}}(\mathbf{z}_2), \ldots, L_{\boldsymbol{\theta}}(\mathbf{z}_n)\Big).$$

Definition 6 (Entropy and Growth)**.** *Let $\mathcal{D}$ and q be as defined above, and $\boldsymbol{\Theta}$ be the set of admissible hypotheses such that, for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\mathbf{z} \in \mathbb{R}^{d+1}$, holds $r_1 \leq L_{\boldsymbol{\theta}}(\mathbf{z}) \leq r_2$, with $r_1, r_2 \in \mathbb{R}$. Assume that $\varepsilon > 0$ is and arbitrary real, and $N(\boldsymbol{\Theta}, \mathcal{D}, \varepsilon)$ denotes the cardinality of the smallest ε-net covering the set $\{q(\boldsymbol{\theta}, \mathcal{D}) | \, \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$. Then, define the following:*

- *Random VC ε-Entropy:*

$$H(\boldsymbol{\Theta}, \mathcal{D}, \varepsilon) := \ln\Big[N(\boldsymbol{\Theta}, \mathcal{D}, \varepsilon)\Big];$$

- *VC ε-Entropy:*

$$H(\boldsymbol{\Theta}, n, \varepsilon) := \mathbb{E}_{\mathbf{Z}}\Big[H(\boldsymbol{\Theta}, \mathcal{D}, \varepsilon)\Big];$$

- *Annealed VC Entropy:*

$$H_{0}(\boldsymbol{\Theta}, n, \varepsilon) := \ln\Big\{\mathbb{E}_{\mathbf{Z}}\Big[N(\boldsymbol{\Theta}, \mathcal{D}, \varepsilon)\Big]\Big\};$$

- *Growth Function:*

$$G(\boldsymbol{\Theta}, n, \varepsilon) := \ln\Big[\sup_{D}\Big\{N(\boldsymbol{\Theta}, \mathcal{D}, \varepsilon)\Big\}\Big].$$

The hypothesis space considered in Definition 6 is general in the sense that it may contain real-valued functions rather than only indicator functions, however, Vapnik and Chervonenkis first developed the theory for the simpler case of indicator functions [Vapnik and Chervonenkis, 1968, 1971] and later generalized it to the version stated above [Vapnik and Chervonenkis, 1981].

The introduction of the VC ε-entropy facilitates the establishment of the necessary and sufficient conditions for the two-sided uniform LLN in the functional space:

Theorem 2 (Generalized Two-Sided Uniform LLN)**.** *Let $\boldsymbol{\Theta}$ be a set of real-valued and bounded dependencies as defined in Definition 6. Then, $\xi_{n} \xrightarrow{p} 0$, as $n \to \infty$, iff*

$$\lim_{n\to\infty} \frac{H(\boldsymbol{\Theta}, n, \varepsilon)}{n} = 0, \quad \forall \varepsilon > 0. \tag{3.4}$$

The generalized two-sided uniform LLN provides a *sufficient* condition for the consistency of the ERM since

$$\mathbb{P}\left\{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\Big[R(\boldsymbol{\theta}) - \hat{R}_{n}(\boldsymbol{\theta})\Big]\right\} \leq \mathbb{P}\left\{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\Big|R(\boldsymbol{\theta}) - \hat{R}_{n}(\boldsymbol{\theta})\Big|\right\}$$

and therefore,

$$\xi_{n} \xrightarrow{p} 0, \quad \text{as } n \to \infty$$

implies that

$$\xi_n^+ \xrightarrow{p} 0, \qquad \text{as } n \to \infty,$$

where the latter has been showed to be a necessary and sufficient condition for the nontrivial consistency of the ERM.

Another important point is that, as we have just seen, grounding the nontrivial consistency of the ERM in the generalized two-sided uniform LLN is obviously too restrictive since the main concern in the ERM is, in fact, the consistency of the minimization problem; more concisely, we are not interested in the fact that whether maximizing the expected risk via the empirical risk is nontrivially consistent. To see this, note that the uniform two-sided convergence can be written as

$$\lim_{n\to\infty} \mathbb{P}\left\{|\xi_n| > \varepsilon\right\} = 0, \quad \forall \varepsilon > 0,$$

where the probability on the left-hand side of the equation might be expressed as

$$\mathbb{P}\left\{|\xi_n| > \varepsilon\right\} = \mathbb{P}\left\{\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left[R(\boldsymbol{\theta}) - \hat{R}_n(\boldsymbol{\theta})\right] > \varepsilon\right) \bigcup \left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left[\hat{R}_n(\boldsymbol{\theta}) - R(\boldsymbol{\theta})\right] > \varepsilon\right)\right\}.$$

But the second event on the right side of the equality above can obviously been violated without affecting the nontrivial consistency of the risk minimization problem. Given this, one should be willing to find a more liberal condition compared to the one expressed in equation (3.4). Hence, as the final step, Vapnik and Chervonenkis [1989] solved the problem of finding the conditions that while being necessary and sufficient specifically for the nontrivial consistency of the minimization part, do not necessarily imply the same thing for the maximization problem.

Let $\boldsymbol{\Theta}$ be the hypothesis space under consideration, which as in previous cases contains only real, and bounded hypotheses. Additionally, assume that there exists another set of hypotheses, say $\boldsymbol{\Theta}^*$, containing measurable functions satisfying the following: For any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ there is a function $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}^*$ such that

$$\begin{aligned} L_{\boldsymbol{\theta}}(\mathbf{z}) - L_{\boldsymbol{\theta}^*}(\mathbf{z}) \geq 0, \quad \forall \mathbf{z} \in \mathbb{R}^{d+1}, \\ \int \left[L_{\boldsymbol{\theta}}(\mathbf{z}) - L_{\boldsymbol{\theta}^*}(\mathbf{z})\right] \, \mathrm{d}F_{\mathbf{Z}}(\mathbf{z}) \leq \delta, \end{aligned} \tag{3.5}$$

where $L_{\boldsymbol{\theta}^*}(\mathbf{z})$ is the loss caused by applying θ^* to $\mathbf{z}$.

Theorem 3 (Necessary and Sufficient Conditions for One-Sided Uniform LLN)**.** *Let* $\boldsymbol{\Theta}$ *be the hypothesis space described in Theorem 1. Then, equation* (3.3) *holds iff for any positive*

δ, η and ε there exists a set of functions $\boldsymbol{\Theta}^*$ with properties (3.5) and such that

$$\lim_{n\to\infty} \frac{H(\boldsymbol{\Theta}^*, n, \varepsilon)}{n} < \eta. \tag{3.6}$$

Note that according to theorem 1 and theorem 3, inequality (3.6) is a necessary and sufficient condition for the ERM to be nontrivially consistent.

So far, the rate of convergence of the empirical risk to the expected one has not been discussed. First, we say that the empirical risk $\hat{R}_n(\boldsymbol{\theta})$ *converges fast* to the true risk $R_0 := \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} R(\boldsymbol{\theta})$ if

$$\mathbb{P}\left\{\hat{R}_n(\boldsymbol{\theta}) - R_0 > \varepsilon\right\} < \exp(-c\varepsilon^2 n),$$

as $n \to \infty$, where c is a positive constant. It turns out that the following condition is *sufficient* for a fast rate of convergence:

$$\lim_{n\to\infty} \frac{H_0(\boldsymbol{\Theta}, n, \varepsilon)}{n} = 0, \qquad \forall \varepsilon > 0. \tag{3.7}$$

However, its *necessity* for a fast rate of convergence is still an open question.

An important point that must be paid attention is that both conditions (3.6) and (3.7) depend on the distribution of the data by definition, while one desires to find general conditions which characterize properties of a learning machine independent from the data distribution. In fact, the following equation, based on the growth function, provides a *distribution-free, necessary and sufficient* condition for nontrivial consistency of the ERM. It also guarantees a fast rate of convergence:

$$\lim_{n\to\infty} \frac{G(\boldsymbol{\Theta}, n, \varepsilon)}{n} = 0, \qquad \forall \varepsilon > 0.$$

Unfortunately, the growth function is hard to calculate in practice. The very important notion of the *VC dimension* is a practical cure to this problem. More precisely, the VC dimension is a basis for providing an upper bound for the growth function, which can be used as measure of complexity for a set of hypothesis. It is important to note that the bound provided by the VC dimension is looser than that given by the growth function, as a result of which the VC dimension is a somehow "less accurate" measure of complexity.

Next, we give the definition of the VC dimension, first, for a set of indicator functions, and then for the general case of a set of real-valued functions.

Definition 7 (VC Dimension of Indicators)**.** *The VC dimension of a set of indicator func-*

tions $Q(\mathbf{z}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, is the maximum number of vectors $\mathbf{z}_1, \ldots, \mathbf{z}_\nu$ that can be separated into two classes in all 2^ν possible ways using the functions of the considered set of indicators. In addition, if any number $\nu \geq 1$ of vectors can be shattered by the considered set of indicators, then the VC dimension of the set of indicators is said to be infinity.

Definition 8 (VC Dimension of Real-Valued Functions)**.** *Let $\boldsymbol{\Theta}$ be the set of admissible hypotheses such that $r_1 \leq Q(\mathbf{z}, \boldsymbol{\theta}) \leq r_2$, with $r_1, r_2 \in \mathbb{R}$. Then, the VC dimension of $\boldsymbol{\Theta}$ is defined as the VC dimension of the indicator functions $\left\{\mathbb{1}_{\geq r}\big(Q(\mathbf{z}, \boldsymbol{\theta})\big) : r \in (r_1, r_2)\right\}$.*

The following results provide non-asymptotic bounds for the empirical risk and will be used later in order to control the capacity of the hypothesis space: Let $\boldsymbol{\Theta}$ be the set of admissible hypotheses such that $r_1 \leq Q(\mathbf{z}, \boldsymbol{\theta}) \leq r_2$, with $r_1, r_2 \in \mathbb{R}$. Then, for any $\theta \in \boldsymbol{\Theta}$

$$\mathbb{P}\left\{\hat{R}_n(\boldsymbol{\theta}) - A \leq R(\boldsymbol{\theta}) \leq \hat{R}_n(\boldsymbol{\theta}) + A\right\} \geq 1 - \eta, \tag{3.8}$$

where

$$A = (r_2 - r_1)\sqrt{\frac{\nu(\ln\frac{2n}{\nu} + 1) - \ln(\frac{\eta}{4})}{n}}, \tag{3.9}$$

and $\nu < \infty$ is the VC dimension of $\boldsymbol{\Theta}$. In addition, if $\hat{\theta}_{(\boldsymbol{\Theta},n)}$ is the minimizer of the empirical risk over $\boldsymbol{\Theta}$, one can verify that

$$\mathbb{P}\left\{R(\hat{\theta}_{(\boldsymbol{\Theta},n)}) - \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} R(\boldsymbol{\theta}) \leq (r_2 - r_1)\sqrt{\frac{-\ln\eta}{2n}} + A\right\} \geq 1 - 2\eta. \tag{3.10}$$

The previous two inequalities (3.8) and (3.10) are related to the following two fundamental questions about the learning ability of a machine:

1. What is the actual risk associated to the hypothesis chosen by minimizing the empirical risk?

2. How close is the actual risk of the chosen hypothesis to the minimum actual risk one can achieve amongst hypothesis in $\boldsymbol{\Theta}$?

Moreover, the aforementioned bounds provide the basis for the method statistical learning theory applies to control the generalization ability of a learning machine.

3.1.5 Structural Risk Minimization

While the SRM is closely related to classical regularization theory in that it provides a more general approach compared to regularization, especially, it provides a better justification for

the use of regularization functionals in data analysis. One issue before the introduction of the SRM was that it was not always straightforward how to justify the application of the classical regularization functionals for a finite set of training examples. These functionals were mainly formulated based on functional analysis arguments, whose concern is clearly not making inference based on empirical data. Consequently, application of these functionals in data analysis relied mostly on asymptotics rather than reflecting finite-data considerations [Evgeniou et al., 2000].

Vapnik's idea can be simplified as follows: When only a finite set of training examples are available, one must search for an optimal hypothesis in a reasonably narrow set of hypotheses since a too "rich" space may contain a function which fits the data perfectly (with zero training error) but performs poorly on new data. As discussed earlier VC theory formalizes these concepts in terms of the VC dimension as a measure of capacity of a set of functions. In addition, it controls the capacity depending on the training data and its size and by applying the bounds explained earlier in section 3.1.4. Put it simply, the bigger the training data, the more complex set of hypotheses can be considered.

Now, to see the SRM in more detail, note that it can be shown that the bounds introduced in section 3.1.4, i.e., expressions (3.8) and (3.10), become arbitrarily narrow as the sample size increases. However, the situation in cases with small sample size is different as explained further.

Before anything, let us clarify that the sample size n will be considered to be small if $\frac{n}{\nu}$ is small, i.e., the ratio of the size over the VC dimension of the machine. As a rule of thumb, we shall regard the sample size is small if this ratio is less than 20.

In the small sample case, the bounds provided by conditions (3.8) and (3.10) do not automatically guarantee the generalization ability of a learning machine. In other words, a small empirical risk does not necessarily result in a small actual risk. Nevertheless, to add to the generalization ability of the learning machine, these bounds suggest not only minimizing the empirical risk but, at the same time, minimizing a combination of the empirical risk and the complexity of the hypothesis space. This leads to the method of *structural risk minimization*, whose main idea is to, based on complexity (capacity), construct a nested structure of hypothesis spaces $\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \ldots, \boldsymbol{\Theta}_k, \ldots$ whose VC dimensions, denoted by $\nu_1, \ldots, \nu_k, \ldots$, satisfy the following:

$$\nu_1 \leq \nu_2 \leq \cdots \leq \nu_k \leq \cdots < \infty$$

Now, for a given set of observations $\mathcal{D} = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$, the SRM chooses a hypothesis $\hat{\theta}_{(\boldsymbol{\Theta}_k, n)}$, which minimizes both (1) the empirical risk over $\boldsymbol{\Theta}_k$, and (2) the summation of the empirical risk and the bound A provided in (3.9).

Let $\hat{R}_{(1,n)}, \hat{R}_{(2,n)}, \ldots, \hat{R}_{(k,n)}, \ldots$ denote the minimum empirical risks achievable over the nested hypothesis spaces, mentioned above, applied to $\mathcal{D}$. Then, it is not hard to see that

$$\hat{R}_{(1,n)} \geq \hat{R}_{(2,n)} \geq \cdots \geq \hat{R}_{(k,n)} \geq \ldots$$

since the larger the hypothesis space, the more flexible it becomes. However, it is well known that, for small n's, this may lead to the problem of overfitting. What the SRM does is controlling the gap between the empirical (training) and expected (true) risk by choosing a reasonable complexity.

Although, the general idea of minimizing the structural risk was first introduced by Vapnik and Chervonenkis [1974a,b] and by using the VC dimension as the measure of complexity, the SRM is a general method and might be utilized with different measures of fit and complexity.

3.1.6 From Glivenko-Cantelli Theorem to Generalized Uniform Convergence

One may summarize the main problem of statistics as *to find an unknown probability measure from observed data*. More precisely, assume that there exists a probability space (Ω, Σ, P), whose probability measure P is *unknown*. The problem of estimating $P(A)$, for any measurable set $A \in \Sigma$, when the measurable space (Ω, Σ) as well as a *limited* number of i.i.d. training examples $\mathcal{D} = \{z_1, z_2, \ldots, z_n\}$ are given, is a fundamental problem in (mathematical) statistics, to which will be referred as the main problem of statistics. Consider the following two approaches to solving this problem:

- *Strong (Complete):* The aim is to estimate $P(A)$ for any measurable subset $A \in \Sigma$, i.e., to completely recover the probability measure;
- *Weak (Partial):* Aims to estimate the probability measure merely for a particular sub-collection of the measurable sets, i.e., to estimate $P(A)$, for $A \in \Sigma'$, where $\Sigma' \subset \Sigma$. In this case, Σ' does *not* need to constitute a σ-algebra.

Investigating the latter problem dates back to 1930s, when a partial solution to the problem of weak estimation was provided by one of the most essential results in mathematical statistics, namely, the *Glivenko-Cantelli Theorem*. It was known from the LLN that the frequency of occurrence of events tends to their probabilities by increasing the amount of observed data, however, none of the LLN's variants cater for the *uniform* convergence of frequencies to the corresponding probabilities. Glivenko [1933] and Cantelli [1933] provided a solution for a certain type of events, i.e., a certain Σ'.

As the theorem was originally proved for the one-dimensional case, we also state it in its original form here. However, generalization to higher finite dimensions can be easily achieved by slight technical modifications. Suppose that Z is a random variable defined on a probability space (Ω, Σ, P) and $\Sigma' \subset \Sigma$ is a set of events of the form

$$\Sigma' := \{\omega \in \Omega \,|\, Z(\omega) \leq t,\, t \in \mathbb{R}\}. \tag{3.11}$$

According to the Glivenko-Cantelli theorem, the frequency of such events approaches their probability asymptotically [Glivenko, 1933, Cantelli, 1933].

Theorem 4 (Glivenko-Cantelli)**.** *Let $Z_1, Z_2, \ldots, Z_n$ be i.i.d. random variables defined over a common probability space and with a common cumulative distribution function F_Z. The empirical distribution function for $Z_1, \ldots, Z_n$ is defined as*

$$\hat{F}_n(t) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(-\infty,t]}(Z_i),$$

where $\mathbb{1}$ is the indicator function. Then,

$$\|\hat{F}_n - F_Z\|_\infty = \sup_{t \in \mathbb{R}} |\hat{F}_n(t) - F_Z(t)| \xrightarrow{a.s.} 0. \tag{3.12}$$

The Glivenko-Cantelli theorem was originally proved for convergence in probability but it holds, in fact, for the almost sure convergence as well, hence equation (3.12). In accordance with the division of the main problem of statistics into weak and strong, the estimation targeted in the Glivenko-Cantelli theorem falls into the weak class of estimations. As one ideally wishes to find a strong estimation tool, it was natural for learning theory to search for viable expansions of the type of convergence achieved by the Glivenko-Cantelli theorem, i.e., the uniform convergence, but for a broader class of empirical measures. Realizing such an expansion gave rise to the introduction of the extremely important *Glivenko-Cantelli class* of functions, which will be discussed shortly.

Before moving onto the introduction of the Glivenko-Cantelli classes, it is worth mentioning that the rate of convergence for the estimator $\hat{F}_n$ in one dimensional case has also been studied separately, for example, by Kolmogorov and Smirnov, among others. We shall skip these results in favour of the more general ones obtained later in Vapnik-Chervonenkis (VC) theory.

To formulate the uniform convergence in a more extensive setting, let us restate the fundamental problem of statistics in the measure-theoretic language: Let (Ω, Σ, P) be a probability

space, [1] whose probability measure P is unknown. The problem is then to estimate P, when the measurable space (Ω, Σ), as well as, sample data $\mathcal{D} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n\}$ are given [Vapnik, 1995]. Note that this original form of the problem does not actually impose any restriction on the form of measurable sets, in contrast with the case of the empirical distribution function, which focuses on a particular type of measurable sets. The need for extension of the Glivenko-Cantelli becomes more apparent by noticing that, besides the huge theoretical appeal this generalization has, considering the empirical distribution function may not be as natural if the sample space Ω consists of objects such as functions, manifolds, etc. This is not a technical concern but rather a conceptual one. The empirical distribution function is a special case of a random measure indexed by Σ' given in equation (3.11). The next concept we would like to introduce is called *empirical measure*, which paves the way for a natural transition from the empirical distribution functions case to the general framework.

Definition 9 (Empirical Measure)**.** *Let $\mathcal{D} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n\}$ be an i.i.d. sample of the random variable $\mathbf{Z}$ from a probability space (Ω, Σ, P) to (S, Σ_s). Suppose that $A \in \Sigma_s$, then the empirical measure of A is defined by*

$$\hat{P}_n(A) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_A(\mathbf{z}_i).$$

Also, if $f : S \to \mathbb{R}^d$ is a measurable function, the empirical measure $\hat{P}_n$ maps f to its empirical mean:

$$\hat{P}_n f := \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{z}_i).$$

Since $\hat{P}_n f$ can be interpreted as the mean of f with respect to the measure $\hat{P}_n$, an alternative notation we will apply interchangeably is $\hat{P}_n f = \int_S f \, \mathrm{d}\hat{P}_n$.

The empirical distribution function is a special case of the empirical measure, where the σ-algebra contains sets of a particular form. The strong LLNs results in both $\hat{P}_n(A) \xrightarrow{a.s.} P(A)$, and $\hat{P}_n f \xrightarrow{a.s.} Pf = \mathbb{E}_P f$, as $n \to \infty$, but as already mentioned, it would be very alluring to be able to have results similar to what the Glivenko-Cantelli theorem provides for the occasion of the empirical measures defined on less restrictive forms of measurable sets. This was, in fact, what the so-called Glivenko-Cantelli classes brought into perspective.

Definition 10 (Glivenko-Cantelli Class)**.** *Let (S, Σ_s, P) be a measure space and $\Sigma' \subseteq \Sigma_s$. Also, let $\mathcal{F}$ denotes a set of measurable functions defined on S. Then,*

[1] The probability space can be replaced by any measure space.

- Σ' *is said to be a Glivenko-Cantelli class of measurable sets if*

$$\|\hat{P}_n - P\|_{\Sigma'} := \sup_{A \in \Sigma'} |\hat{P}_n(A) - P(A)| \xrightarrow{a.s.} 0, \qquad \textit{and}$$

- $\mathcal{F}$ *is said to be a Glivenko-Cantelli class of measurable functions if*

$$\|\hat{P}_n - P\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |\hat{P}_n f - Pf| \xrightarrow{a.s.} 0,$$

as $n \to \infty$.

Therefore, one of the objectives of theory can be summarized as to find a full characterization of the Glivenko-Cantelli classes of functions. This characterization was achieved by providing the necessary and sufficient conditions for a class of functions or sets to be Glivenko-Cantelli.

3.2 A Discussion on Ill-Posed Problems and Inductive Bias

According to Hadamard, a *well-posed* mathematical problem must have the following properties: (a) existence of a solution, (b) uniqueness of the solution, and (c) continuity (stability) of the solution with respect to the initial conditions. Any problem not satisfying any of the aforementioned properties is called *ill-posed*.

There is a broad category of scientific and mathematical problems, called *inverse problems*, whose main goal is to find some model's parameters by looking at a finite set of observed instances of the model [Tarantola, 2005]. Apparently, the general problem of learning from examples falls into this category. Moreover, a considerable amount of inverse problems, including the learning problem, are typical examples of ill-posed problems in Hadamard's sense. Particularly, learning form examples violates Hadamard's condition on the uniqueness of a solution because it involves inducing a parametric function $h_{\boldsymbol{\theta}}$, which describes the relation between the variables x and y, from a finite set of points $\mathcal{D} = \{(\mathbf{x}_i, y_i) | i = 1, \ldots, n; n \in \mathbb{N}\}$, contaminated by some noise. Clearly, there might exist multiple (usually infinite) values of the parameter which fit the data reasonably well. Therefore, the general problem of learning, in the sense explained, is not solvable.

Classical methods to approach ill-posed problems usually involve introducing some additional constraints to the initial problem either by (i) restricting the parameter space (hypothesis space), or (ii) targeting an "optimal" solution rather than a unique universal one, or alternatively by a combination of both. The well-known regularization methods are examples of classical methods created for dealing with ill-posed problems.

Apart form the validity of possible justifications for adding this additional information to the analysis, added constraints lead to imposing a certain type of bias to the model which is referred to as *inductive* or *learning bias*. An inductive bias is essentially an auxiliary assumption without which the learning problem cannot be solved uniquely. It will not be exaggerating to claim that the majority of inferential methods, applied in machine learning and statistics, suffer (more precisely, benefit) from one or more inductive biases [Mitchell, 1980, Gordon and desJardins, 1995].

The next section discusses a specific method used in statistical learning theory to approach the ill-posed learning problem. But before moving further, and in order to grasp a sense of inductive bias, let us give some examples of inductive biases in popular machine learning methods:

- The simplest example is the ordinary linear regression problem, where the hypothesis space is reduced to the set of linear functions rather than allowing the hypothesis space to include any possible function.
- The naive assumption of conditional independence in another example of an inductive bias in models like the naive Bayes classifier.
- Classifiers that classify objects based on maximizing a separating margin, such as the *support vector machines*, are also biased in that they assume the a wider gap between classes provides better classification.
- The famous *Occam's razor*, which favors less complex models, in fact, is the next example of a learning bias. Perhaps, this is one of the most frequently inductive biases met in a vast variety of scientific fields. Indeed, the rational behind the *variable selection* methods is a special case of the Occam's razor principle.
- *Instance-based* learning algorithms, which apply notions like *similarity*, *association* or *distance*, additionally, assume that closer or more similar objects are more likely to belong to the same class. A known example is the K*-nearest neighbors* algorithm. Note that a similar concept is used in the K*-means clustering* algorithm too.

Since studying the inductive bias types is not in the scope of this note, we will not extend the above list further as it can get too exhaustive. More examples might be found on Wikipedia and Mitchell [1980].

3.3 Time-to-Event Data Analysis and Related Issues

As stated in Chapter 1, analysis of Time-to-event, survival time, or lifetime arises often and in numerous fields, which makes it necessary to study and to develop suitable techniques that can be applied to the analysis of such data. Especially, due to the specific properties of time-to-event data, usually, these techniques require to be designed particularly for dealing with these certain peculiarities. In this section, we introduce some of the important concepts, procedures, and issues related to the analysis of survival data and, especially, explain the weakly-supervised setting we consider in the following chapters.

3.3.1 Basic Concepts

First of all, let us define the main variable of interest, i.e., the time-to-event. Note that we will use *time-to-event*, *survival time*, *lifetime*, and *failure time* interchangeably, throughout this thesis. Time-to-event is defined as the amount of time elapsed from the occurrence of an *initiating event* E_1 until that of a second event E_2, called the *terminating event*. Both events are pre-defined. E_1 might be birth, or onset of a disease, e.g., while the corresponding terminating events, E_2, could be defined as death, or the recurrence of a disease, respectively. In the rest of the thesis, Y will be used to denote the survival time.

Now, having the primary variable of interest defined, in the coming paragraphs, we will briefly introduce several basic concepts which are of central importance in the analysis of time-to-event data.

Let Z be a positive random variable and z be an arbitrary point in its range. By F_Z we denote the CDF of Z, i.e. $F_Z(z) = \mathbb{P}(Z \leq z)$. If Z represents the failure time of a certain object, then the *survival function* will be denoted by S_Z and is define as

$$S_Z(z) := \mathbb{P}(Z > z) = 1 - F_Z(z).$$

The *hazard function*, denoted by $h_Z(z)$, is the conditional probability of failure, given survival up to time z. That is,

$$h_Z(z) := \lim_{\Delta z \to 0} \frac{\mathbb{P}(z \leq Z < z + \Delta z | Z \geq z)}{\Delta z}.$$

Consequently, the *cumulative hazard function* may be achieved by integrating the hazard

over the time period between 0 and z:

$$H_{Z}(z) := \int_{0}^{z} h_{Z}(u)\,\mathrm{d}u.$$

It is easy to see that

$$h_{Z}(z) = \frac{f_{Z}(z)}{S_{Z}(z)} = -\frac{\mathrm{d}}{\mathrm{d}z}\left[\ln S_{Z}(z)\right],$$

where f_{Z} is the probability distribution function (PDF) of Z.Different basic types of hazard can be distinguished based on how it changes over time.

3.3.2 Sampling Procedure and Its Consequences

The so-called *follow-up* studies are among the most popular study designs being used in survival analysis. In fact, conducting follow-up studies on a *randomly* selected sample of the *target* population is considered as the gold standard in related areas. Nonetheless, there exist situations where keeping up with this standard is extremely hard or even impossible. The next concept to be explained is closely related to such situations. In the context of time-to-event, the target population, mentioned above, is referred to as the *incident population*. That is, the population whose individuals have not experienced the initiating event E_{1}, and consequently E_{2}, before the study begins. The *recruitment time*, or simply the *recruitment* will be used to refer to the start of a study. Different factors may, indeed, prevent one from sampling directly from the incident population. To name a few, consider logistic restrictions, for instance, or the financial burdens this ideal design imposes on the available material resources.

Instead, a feasible alternative in such cases is to conduct a *cross-sectional prevalent cohort study*, where one recruits *prevalent* cases rather than incident ones [Huang and Wang, 1995, Wang, 1991, Wang et al., 1993]. More concisely, the recruited subjects are being chosen from the portion of the incident population who have already experienced E_{1} but not E_{2}. When the interest lies in estimating the lifespan between the initiating and the terminating events, subjects may be followed prospectively either until the terminating event happens or until they are lost to follow-up, whichever occurs first.

Such a sampling scheme gives rise to two types of incompleteness:

First, the response variable Y, i.e., the time interval between E_{1} and E_{2}, might be, exactly, observed only for a subset of the recruited subjects. For the rest of the sample, we only know that the terminating event has *not* occurred up until a specific moment in time. This specific moment is called the *censoring time*. There are multiple sorts of censoring, which will be

explained, in detail, later in this section.

Second, it is well known that the prevalent cases have, on average, longer lifetimes since longer survivors are more prone to be recruited to the study. This leads to a phenomenon referred to as *truncation*, i.e., some of the individuals in the incident population have a smaller chance to be selected by this specific sampling method.

The term *weakly-supervised learning* in the thesis' title, in fact, points to situations where learning is happening in presence of incomplete data. This falls into a subcategory of the more general *supervised* learning since the response variable Y is clearly defined, i.e., the survival time, and also the value of Y is available, although partially. Note that a prevalent cohort comprises a non-random sample that is not representative of the target incident population. In the following paragraphs, we will discuss typical time-to-event data characteristics in more detail. Particularly, we focus on the deeper aspects of incompleteness, including censoring, truncation, and their types.

3.3.3 Essential Data Characteristics

In the following sections, we provide a general introduction to some data subtleties that are most frequently met in survival data. Note that besides these specific peculiarities, survival data can be vulnerable to other general data pathologies as well.

Non-Normality of Lifetime

One of the very first distinctions of time-to-event data is that most of the time the *data are not normally distributed* and instead have a sort of asymmetric distributions reflecting the fact that the event of interest, i.e., the terminating event, often tends to occur, for example, at earlier stages of the experiment or vice versa. An example is the infant mortality rate as newborns are normally more susceptible to health problems compared to children of a few weeks age.

Incompleteness and Bias

A principal objective of any data analysis is to generalize the information, obtained from a sample of a population, to the whole population by means of induction. Hence, data are one of the main tools based on which statistical inference is made. This fact makes it vitally important to make sure, in the first place, that the data in hand satisfy necessary conditions for making a sound inference. It should not be difficult to figure out how undesirable qualities

of data may threaten the soundness of the inference if being ignored or not taken into account properly. The following example highlights this fact.

A well-known dilemma, called *the falling cat*, refers to a 1987 study, where the collected data suggested a controversial outcome as follows. 132 cats that had fallen out of high-rise windows and were brought to a veterinary hospital were examined for related injuries. Surprisingly, analyzing the data showed that falling from higher stories were associated with less severe injuries. However, some researchers argue that this counter-intuitive finding could have been the result of a bias called the *survivorship bias* [Whitney and Mehlhaff, 1987].

Incompleteness results in the sampled data being *unrepresentative*. This is the case in the setting considered in this work: Truncation, systematically, introduces bias to the collected data and, therefore, make them unrepresentative of the whole population. It is worth mentioning that, in practice, there exists a vast range of sources that might be responsible for the incompleteness of data. Also, biasedness and missingness can happen at different phases of data collection or even analysis itself. Sometimes, bias is an inevitable result of employing a certain method of sampling (like in the situation of our interest), while other times, it could be a byproduct of the intrinsic nature of the stochastic phenomenon being studied. It would be insightful to see a few examples, where data are contaminated by some sort of bias or incompleteness. Next, we focus on two specific types of incompleteness that determine the setting of our interest.

Incompleteness Due to Censoring

Before giving the exact definition of censoring, let us begin with a simple example, to best understand it. Consider a scale that is able to weigh objects up to a maximum of 3 tons. If one places an object weighing more than 3 tons on it, the scale still indicates the maximum measurable value, i.e., 3 tons. Hence, while the exact weight of this object cannot be known using this scale, it still provides partial information about it, i.e., the smallest upper bound of its weight.

As in the bathroom scale example, censoring refers to situations, where the information about the realized value of a random variable is available only partially. As an example, assume that we are interested in the *relapse-free survival time* corresponding to a certain treatment of a disease. This is the amount of time from receiving the treatment to the recurrence of the disease. However, in survival data, often, not all the subjects have experienced the terminating event due to either of the following reasons: (1) A patient has not yet experienced the recurrence by the end of the study; (2) A patient had been lost to follow-up during the study; or (3) A patient has exited the study for a reason irrelevant to the study. The aforementioned

scenario comprises one of the multiple types of a phenomenon, called *censoring*. Technically, three types of censoring happen in survival data, which are defined as follows.

Definition 11 (Censoring)**.** *Let Y be a real-valued random variable and $a, b \in \mathbb{R} \cup \{-\infty, \infty\}$, and $a \leq b$. Also, let y be a realization of Y, whose exact value is unknown. Then, y is said to be (1)* interval censored *if $y \in (a, b)$; (2)* left censored *if y is interval censored and $a = -\infty$; and (3)* right censored *if y is interval censored and $b = \infty$.*

The most common type of censoring in survival data is the right censoring; the relapse-free survival time example, given earlier in this subsection, is also an occasion of right censoring. For an example of the left censoring, imagine that in the previous example patients are examined for recurrence of the disease 6 months after the administration of the first dose of the treatment. Those subjects for whom a relapse is detected in the exam, the exact disease-free time is unknown and the only information available in this setting is that the value of interest is less than 6 months. And finally, if a second examination is performed after a year, then those subjects who were disease-free in the first exam but otherwise in the second exam are examples of interval-censored data. Similarly, a disease-free individual in the first exam who has been lost to follow-up before the second exam is also regarded as an interval-censored case.

Moreover, right censoring itself is divided into two sub-types:

1. *Type I:* It happens when at a *predetermined time* all the remaining subjects in the study, i.e., those who have not experienced the failure event yet, are censored.

2. *Type II:* It occurs when a *predetermined number of failures* is reached, regardless of the time at which the failures have happened the rest of the participants are censored. Note that a failure means experiencing the terminating event.

Yet, another point to be taken into account in censored data is the so-called *informativeness* of the censoring mechanism. Censoring is called *random* or *non-informative* if censoring and failure times are independent from each other. Notice that there is a close connection between the randomness of censoring and that of the missingness, in general.

Most attention, especially, in survival analysis and related domains, has been paid to statistical methods dealing with right-censored data. Nonetheless, there have, also, been methods developed for treating left and interval censoring. (See Hosmer et al. [2008], for example.)

Throughout the remainder of this thesis, right-censored data are of our interest. Note that when data are subject to right censoring, one only observes the minimum of the censoring

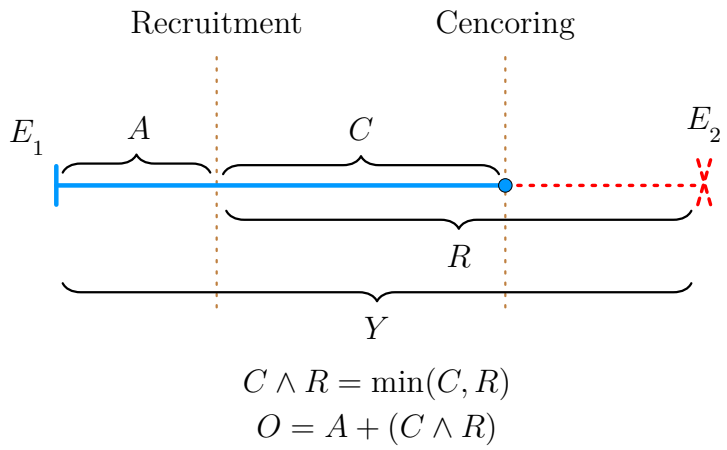

Figure 3.3 **Available Information on Each Subject.** Here, a right-censored subject from a (potentially left-truncated) sample is illustrated. Values displayed are as follows: E_1, E_2 are the initiating (onset) and terminating (failure) events; Y is the time-to-event; A shows the current lifetime, while R is the residual lifetime; and lastly, C indicates the residual censoring time. O is the observed value of the "response", which is either the total censoring or failure time. In case of left-truncated data $\tilde{Y}, \tilde{A}$, and $\tilde{R}$ would be used.

and failure time, i.e., $\min(C, R) := C \wedge R$. C is called the *residual censoring time* and is the time interval from the recruitment until the moment the potential censoring of the subject happens. The *overall censoring time*, however, refers to the time interval between the initiating event E_1 and the potential moment of censoring and is denoted by C'. Finally, R is the *residual lifetime* or *forward recurrence time*, i.e., the time interval elapsed from the recruitment of a subject up to its failure E_2. Notice that the survival time Y, which is the main variable of interest is the whole interval between E_1 and E_2. (See Figure 3.3.)

Weighted Distributions, Length Bias and Truncation

Sampling or selection bias is a phenomenon, closely, related to the so-called *weighted distributions*. Let Z be a random variable with density f_Z. In standard situations, i.e., where the sampling procedure is not biased, inference about f_Z is made directly through the drawn sample as the observations are distributed according to the same distribution as the target population, i.e., f_Z. However, there are cases, in practice, where the data are observed from a distribution proportional to $w(z)f_Z(z)$, where $w(z) \geq 0$ is a weight function defined on the range of Z. Observing such practical situations, in fact, motivated the definition and investigation of the weighted distributions. One of the earliest considerations of weighted distributions took place in a study by Fisher [1934], as he investigated the impacts of the methods of ascertainment on the sampling frequencies of random events. Later Rao [1965, 1985] studied the distribution of random variables as they were observed (being sampled) and

generalized the concept of a weighted distribution by extending it from observed frequencies to distributions. We define the weighted density of the observations as

$$f_{_w}(z) = \frac{w(z) f_{_Z}(z)}{\mathbb{E}_{_Z}[w(Z)]},$$

provided that $\mathbb{E}_{_Z}[w(Z)] < \infty$. The above definition may be extended to the case of discrete random variables easily. An interesting example is the phenomenon of *publication bias* in meta-analysis. This is a bias that happens due to a tendency of "not to report the details of nonsignificant results" when combining findings in multiple studies on the same issue. That is, researchers, oftentimes, leave the details of statistical analyses unreported when mean differences obtained are not statistically significant. In other words, detailed statistical results become published proportionally to the significance of the results, where the weight function can be expressed as $w(z) = \mathbb{1}\{|z| \geq 1.96\}$. The failure to report nonsignificant results is considered as a type of *prejudice against the null hypothesis*. (For results concerning the publication bias in meta-analysis, see Hedges [1992], Iyengar and Greenhouse [1988]).

A special case of weighted distributions is of great interest in survival analysis and reliability theory for the following reason: As mentioned earlier, there are numerous situations where keeping up with standard sampling schemes in observational studies is not the preferred one due to different reasons. When, instead, prevalent cases are recruited to a study, the chance of selecting an individual becomes proportional to the individual's survival time, which is the variable of interest. Technically speaking, we are sampling cases according to a weighted distribution with $w(z) = I(z) = z$, where I represents the identity function. Put it otherwise, the random variable Z is observed with a probability proportional to its size. This selection bias is called *size-bias*. In survival analysis however, *length bias* is the favored term pointing to the variable of interest, i.e., the length of survival time. Accordingly, replacing the weight function w with the identity function in the equation above gives the density of the length-biased Z, denoted by $f_{_{\text{LB}}}$, as

$$f_{_{\text{LB}}}(z) = \frac{z\, f_{_Z}(z)}{\mathbb{E}(Z)}. \tag{3.13}$$

To explain the relation between the length bias and the specific sampling method being used, let us define the sampling procedure in a formal language. Suppose that Z is a random variable with density $f_{_Z}$. Formally, one can consider *sampling* as a procedure during which some of the already realized values of the random variable Z are being selected or observed. This interpretation enables us to formalize the sampling procedure as a separate random variable defined on the same sample space and σ-algebra as Z's but with a different probability measure. Accordingly, let $\zeta_{_Z}$ be a random variable representing a *sampling procedure* from Z.

Then, the event of a certain realization z being sampled by ζ_Z is expressed by $\{\zeta_Z = z\}$. This way, one can define a length-biased sampling procedure: ζ_Z is said to be *length biased* if

$$\mathbb{P}\{\zeta_Z = z\} \propto z.$$

Further, we will use a tilde sign to denote a length-biased variable. Accordingly, $\tilde{Y}$ will be applied to denote length-biased survival time.

The next important concept to be defined here is called *truncation*. Recall that an incident population consists of individuals who experienced the incidence of the condition of interest sometime within a specified interval of time. However, they may or may not have had the terminating event or failure. In contrast, a prevalent population refers to individuals who, at a specified moment, have experienced the incidence of the condition but the defined failure has not occurred for them.

Note that because the incident population is defined independently from the occurrence of failure, the prevalent population is a subset of the incident population. This simple fact, makes it thoroughly reasonable to aim at the incident population for drawing a sample instead of the prevalent population if the goal is to make inference about the survival time in connection with the considered condition. But as mentioned before this is not always possible or preferable out of different reasons, such as budget shortage or ethical concerns.

As explained earlier sampling from the prevalent population is a popular alternative. Particularly, this is a major scenario in many observational studies like the popular cross-sectional cohort studies. This, obviously, makes the sample biased by excluding part of the population from the sample. In fact, this selection scheme ignores individuals who have already experienced both events before the enrollment, e.g., those who had experienced the disease sometime in the past and died subsequently as a result of the illness before the enrollment time. It means that to have the chance to be enrolled in the study, an individual has to, at least, survive up to the recruitment time. This leads us to the definition of truncation in general, and the definition of left truncation, in particular.

Definition 12 (Truncation)**.** *Let Y be real-valued random variable and ζ_Y a sampling scheme from realizations of Y, where selection of a certain y by ζ_Y is denoted by $\{\zeta_Y = y\}$. In addition, suppose that $a, b \in \mathbb{R} \cup \{-\infty, \infty\}$, and $a \leq b$. Then, we say that ζ_Y leads to*

1. interval truncation *or* truncation from below and above *if* $\{\zeta_Y = y\} \implies y \in (a, b)$*;*
2. left truncation *or* truncation from below *if it leads to interval truncation and* $b = \infty$*;*
3. right truncation *or* truncation from above *if it leads to interval truncation and* $a = -\infty$*.*

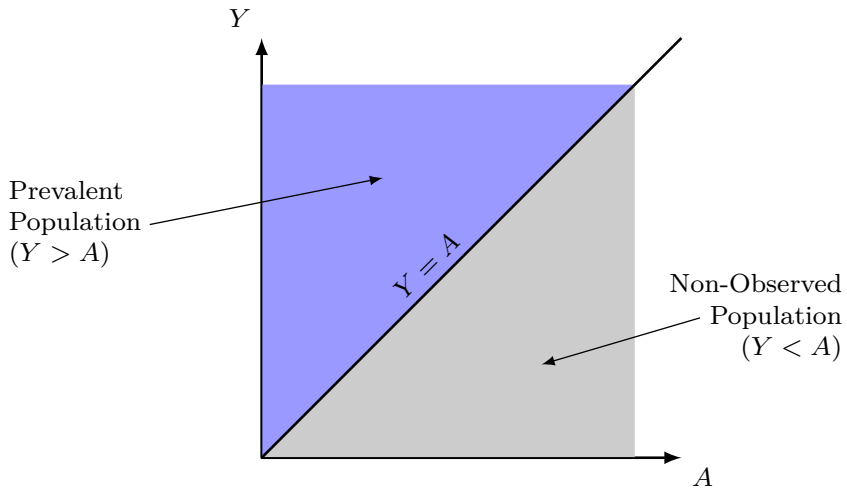

Figure 3.4 **Incident vs Prevalent Populations.** Here, the prevalent population is depicted versus the incident one. The purple upper triangle displays the prevalent population, while the grey lower triangle represents individuals having experienced both initiating and terminating events. The union of these two areas, i.e., the whole square, makes the susceptible (incident) population. Y denotes the lifetime and A, the truncation time.

Further, we will refer to the time interval between the initiating event and the recruitment by the term *truncation time*, *current lifetime*, or *backward recurrence time* and will denote it by A.

According to the given definitions, one may see that the sampling design mentioned in this section, generates left-truncated samples. Nevertheless, one must be aware that a sample being left truncated does not necessarily imply its length biasedness. More precisely, left truncation does not result in the selection density stated in equation (3.13). To generate length biasedness out of left truncation, it must be accompanied by an extra condition, called *stationarity*. Stationarity requires the truncation time to be distributed uniformly among the incident population. The union of the left truncation and stationarity assumption leads to the probability of observing a certain value y (of random variable Y) is proportional to its size, i.e., $f_{\tilde{Y}}(y) \propto y f_{Y}(y)$. This union constitutes a very important assumption throughout the current thesis. Hence, the length bias is the union of the uniformity of A together with $Y \geq A$ and will be denoted by a *tilde* sign. As the main interest of this chapter involves studying length-biased samples, in the rest of the chapter stationarity is always assumed to hold when left truncation is considered. (More details about the relation between the distribution of onsets and the bias induced by left truncation has been described by Brookmeyer and Gail [1987]. (see Figure 3.4).

It is important to note that truncation is the result of the sampling procedure, whereas censoring is not. In fact, truncation is related to *ascertainment bias*, *sampling bias*, *survivor-*

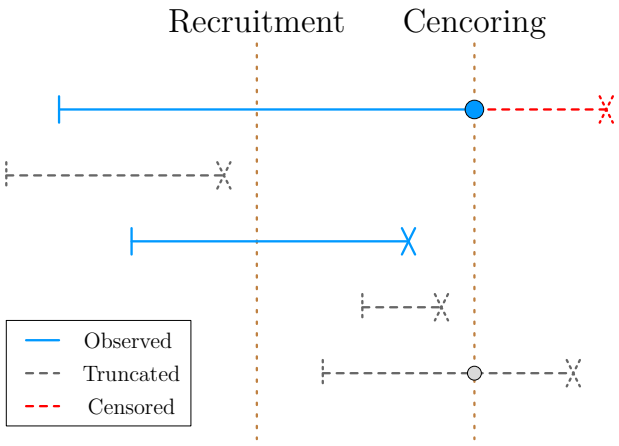

Figure 3.5 **Left Truncation and Right Censoring.** Incompleteness of data due to left truncation and right censoring is depicted here. The dashed grey subjects are excluded from the study as they do not belong to the prevalent cohort.

ship/survival bias, *exclusion bias*, and *caveman effect*. While censoring causes a partial loss of information on some subjects enrolled in the study, truncation excludes some subjects from being sampled completely causing a total loss of information on those subjects, even their existence. One must distinguish between a bias that causes missing data and a bias that is the result of missing data.

3.3.4 Impact of Length Bias on Covariates

Inclusion of covariates into the analysis of data, collected through cross-sectional sampling, adds new concerns that must be paid close attention. With the exception of few studies the covariate-related issues in left-truncated data have been mostly skipped in the literature. Data collection through cross-sectional prevalent cohort sampling, in addition to the introduction of length bias to the response variable, imposes bias on the sampling distribution of the covariates as well and, as we will see later, failure to include this information into the analysis may cause highly misleading results. Particularly, in most conventional regression analysis, the tradition of conditioning on the observed values of covariates results in missing out on the information provided by the sampling distribution of the covariates. Clearly, this is an issue if the covariate distribution is indeed informative. The reason why length bias affects the sampling distribution of covariates may be explained as follows. Since the discussed sampling design tends to select the items with larger values of survival time it is natural to expect an over-representation of the covariate values which are associated with longer survivors. Hence, the induced covariate bias in the prevalent cohort is another consequence of the sampling mechanism. Distribution of the biased covariates will be discussed in detail in

later sections.

3.3.5 Summarizing the Notation and Terminology

Here, we complete and summarize the specific notation adopted for the analysis of LBRC-C data in the rest of the present thesis.

- Uppercase letters denote one-dimensional random variables, while bold uppercase denotes random vectors (or matrices).
- Lowercase bold and regular letters are reserved for realizations of random vectors and random variables, respectively.
- The sample or training data is represented by $\mathcal{D}$ throughout the thesis.
- Y denotes the response variable, which is time-to-event, survival time or lifetime in the rest of this work unless otherwise specified.
- Length-biased variables will be marked with a tilde, e.g., $\tilde{Y}$ refers to the length-biased survival time. Naturally, we assume that $Y, \tilde{Y} \geq 0$.
- $\mathbf{X} = (X_1, X_2, \ldots, X_d)$, with $d \geq 1$, is a vector of covariates.
- Biased variables (not length-biased) will be represented by an asterisk. Therefore, $\overset{*}{\mathbf{X}}$ is applied to denote a biased vector of covariates.
- Bold $\mathbf{Z}$ is used to show the vector of covariates and the response together, i.e., $\mathbf{Z} = (\mathbf{X}, Y)$. Similarly, $\overset{*}{\mathbf{Z}} = (\overset{*}{\mathbf{X}}, \tilde{Y})$.
- For realization of the random vectors or variables we do not use any tilde or asterisk, i.e., $\tilde{Y} = y$ or $\overset{*}{\mathbf{X}} = \mathbf{x}$.
- When applicable, regression coefficients are denoted by $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_d)$ or its transpose. Nevertheless, $\boldsymbol{\theta}$ is used to indicate the vector of all parameters to be estimated, including the regression ones.
- A is the current lifetime, i.e., the time interval between the initiating event and sampling time.
- Residual lifetime or backward recurrence time is denoted by R. Therefore, $Y = A + R$.
- C_i is the censoring time, which is the time elapsed from the sampling of the subject until its possible censoring.

- One may observe only $R \wedge C = \min(R, C)$ due to possible censoring.
- Additionally, we define $O := A + (R \wedge C)$.
- The failure indicator δ is defined to be a Bernoulli random variable indicating whether a subject has failed or censored; that is, $\delta = \mathbb{1}_{\{R \leq C\}}$.
- The relevant information per each subject can be represented as a random vector, each entry of which being a feature. For brevity, we assume that any sample $\mathcal{D}$ consists of subjects $\mathbf{S}_i$, $i = 1, 2, \ldots, n$, where $\mathbf{S}_i$ is a vector encompassing all features associated with subject i. Therefore, a sample dataset with n subjects, in its most general form, i.e., without specifying the features is represented as $\mathcal{D} = \{\mathbf{S}_i : i = 1, \ldots, n\}$. If data includes the time-to-event for each subject, then $\mathbf{S}_i = (Y_i) = Y_i$; or in the context of our interest we have that $\mathbf{S}_i = (\overset{*}{\mathbf{X}}_i, \tilde{A}_i, \tilde{R}_i \wedge C_i, \delta_i)$, i.e.,

$$\mathcal{D} = \left\{ (\overset{*}{\mathbf{X}}_i, \tilde{A}_i, \tilde{R}_i \wedge C_i, \delta_i) : i = 1, \ldots, n \right\},$$

  where $\overset{*}{\mathbf{X}}_i = (\overset{*}{X}_{i_1}, \ldots, \overset{*}{X}_{i_d})$ (see Figure 3.4).

3.3.6 Classical Approaches to Time-To-Event Regression

Apart from the survival time itself, in many situations, one might be interested in how a set of covariates (inputs) affect the survival rate. Regression analysis of time-to-event data allows one to investigate the impact of a set of explanatory variables on the survival time when analyzing survival data. In what follows, we briefly introduce the most commonly applied regression models of lifetime data in the classical framework of statistical inference, namely, the *CPH*, and the *accelerated failure time (AFT)* models.

Cox Proportional-Hazards Model

The famous CPH model, introduced by Cox [1972], is a multiple regression model, which, like any other multiple regression model, provides the possibility of studying the simultaneous effects of several risk factors (covariates) on the hazard rate at different points of time. The core assumption in the CPH model states that the effect of each covariate on the hazard rate at time t is to increase or decrease the hazard by some constant, i.e.,

$$h(t|\,\mathbf{X}; \boldsymbol{\beta}) = h_0(t) \exp\langle \mathbf{X}, \boldsymbol{\beta} \rangle\,, \tag{3.14}$$

where h_0 is unknown and is called the *baseline hazard*, i.e., the hazard rate when all covariates equal zero:

$$h_0(t) := h(t|\,\mathbf{X} = \mathbf{0}).$$

$\boldsymbol{\beta} = (\beta_0, \ldots, \beta_d)$ is a vector of parameters, commonly called regression coefficients, describing the relative influence or risk associated with each covariate (or factor). Widely speaking, $\exp\langle \mathbf{X}, \boldsymbol{\beta}\rangle$ in equation (3.14) (sometimes called the *relative hazard function*) could be replaced by any suitable known function of the covariates and their coefficients [Cox, 1972]. The proportionality suggests that if a subject's risk of failure is r times of the failure risk of another subject, then at all points in time, the risk ratio remains the same, i.e., r. The proportional hazards model is a semi-parametric model as the baseline hazard is left completely unspecified. In other words the baseline hazard h_0 is an infinite-dimensional parameter, whilst $\boldsymbol{\beta}$ is finite-dimensional. In practice, the baseline hazard can be estimated by the *Breslow's estimator* [Breslow, 1975].

By looking at equation (3.14) one could see that an obvious advantage of the CPH model is, undoubtedly, the straightforward interpretation it provides due to the simple form of the function chosen to express the impact of the covariates on the hazard. More precisely, the effect size of each covariate could be easily verified by the magnitude of its coefficient. A coefficient greater than zero indicates a positive association between the corresponding covariate and the hazard rate, while a negative coefficient is an indicator of a negative association.

Sometimes, researchers, instead of the coefficient β_j, $1 \leq j \leq d$, itself, use the *hazard ratio* to determine the effect of covariates on the hazard or survival. Hazard ratios are defined as $\exp(\beta_j)$. In cancer studies, for example, a covariate with a hazard ratio greater than 1 (equivalently, a positive coefficient) is called a *bad prognostic factor*, whereas a covariate with a hazard ratio less than 1 (equivalently, a negative coefficient) is called a *good prognostic factor*.

Similarly, as in any model-based analysis of data, it is important to validate the adequacy of the model in terms of describing the dependence between the failure time and covariates. One of the things to be examined is the validity of the assumptions made by the CPH model. Fortunately, the CPH model makes a fairly minimal set of assumptions including the aforementioned proportional-hazards, linearity of the relation between the hazard and covariates, and influential observations. Commonly applied residuals tests are (i) *Schoenfeld* residuals to check the proportional hazards assumption [Schoenfeld, 1980, 1982], (ii) *martingale* residuals to assess linearity, and (iii) *deviance* residuals to examine influential observations [Grambsch and Therneau, 1994].

Accelerated Failure Time

The AFT (or accelerated life) time model is, also, a regression models that explores the relation between potential risk factors and the failure time. Similarly, as for the CPH models, the AFT model is based on a set of assumptions, some of which are analogous to those of the CPH model such as linearity. Nonetheless, the primary distinction lies in the assumed effect of the covariates on the variable of interest, i.e., the survival time. The AFT assumes that the effect of the risk factors on the survival time is to accelerate (or inversely, to decelerate) the survival time by some constant. Therefore,

$$S(t|\,\mathbf{X};\boldsymbol{\beta}) = S_{_0}\left(t\,\exp\left\{-\langle\mathbf{X},\boldsymbol{\beta}\rangle\right\}|\,\mathbf{X};\boldsymbol{\beta}\right), \tag{3.15}$$

where $S_{_0}(t)$ is the probability that a reference subject, i.e., a subject for whom $\mathbf{X}=\mathbf{0}$, will be alive at time t. By analogy, we may call $S_{_0}$ the *baseline survival function*. An equivalent way of formulating an AFT model is by the following ordinary linear regression model for in terms of the log-survival time:

$$U = \log Y = \langle\mathbf{X},\boldsymbol{\beta}\rangle + \sigma\varepsilon, \tag{3.16}$$

where ε is a suitable error term, and σ is a constant multiplier that modifies the noise's significance and is independent of $\mathbf{X}$. In fact, different error distributions lead to different baseline survival distributions. In (3.16), the term $\sigma\varepsilon$ gives the baseline distribution of the lifetime, i.e., when covariates are all zero. To see the effect of the covariates, it is enough to exponentiate both sides of equation (3.16) which gives

$$Y = Y_{_0}\,\exp\{\langle\mathbf{X},\boldsymbol{\beta}\rangle\},$$

with $U_{_0} := \exp(\sigma\varepsilon)$. Now, we have that

$$\begin{aligned} S(t|\mathbf{X};\boldsymbol{\beta}) &= \mathbb{P}\{Y > t|\,\mathbf{X};\boldsymbol{\beta}\} \\ &= \mathbb{P}\{Y_{_0}\exp\{\langle\mathbf{X},\boldsymbol{\beta}\rangle\} > t|\,\mathbf{X};\boldsymbol{\beta}\} \\ &= \mathbb{P}\{Y_{_0} > t\exp\{-\langle\mathbf{X},\boldsymbol{\beta}\rangle\}|\,\mathbf{X};\boldsymbol{\beta})\} = S_{_0}(t\exp\{-\langle\mathbf{X},\boldsymbol{\beta}\rangle\}|\,\mathbf{X};\boldsymbol{\beta}), \end{aligned}$$

which is equation (3.15). This means that the probability of an individual with covariate $\mathbf{X}$ to survive up to time t equals the probability that a reference subject (with $\mathbf{X}=\mathbf{0}$) will be alive at time $t\exp\{-\langle\mathbf{X},\boldsymbol{\beta}\rangle\}$. In other words, time passing has accelerated by a factor of $\exp\{-\langle\mathbf{X},\boldsymbol{\beta}\rangle\}$. While everything happens as $\exp\{-\langle\mathbf{X},\boldsymbol{\beta}\rangle\}$ times faster, unlike the CPH,

this does not mean that the hazard function is always $\exp\{-\langle \mathbf{X}, \boldsymbol{\beta} \rangle\}$ times as high.

The AFT model is usually applied in a fully parametrized manner, that is, the distribution of the baseline survival time is specified completely. However, it is possible to take a semi-parametric approach as well, for instance, see Buckley and James [1979].

CHAPTER 4 FOUNDATIONS OF LEARNING FROM INCOMPLETE DATA

Earlier in Chapter 3, the importance of investigating the learning problem in the context of sampling bias and incompleteness due to censoring has been discussed. It was, also, mentioned that this importance stems mainly from the inevitable constraints typically imposed by limitations of time or other resources. In particular, we have seen that the common use of the cross-sectional, prevalent-cohort design with follow-up leads to left truncation and right censoring. This is quite a frequently encountered setting in practice. To see some examples, check Huang and Wang [1995], Wang [1991], Wang et al. [1993], among others.

Considering the importance of the problem, the present chapter is devoted to studying a few problems that are connected to the aforementioned setting and learning from LBRC-C data. It is important to note that, although we emphasize, particularly, the length bias throughout the discussion, most of the results obtained in this chapter could be easily extended to other types of bias, so long as the bias' structure is known. The chapter consists of three primary sections discussing the following problems, independently:

1. Estimation of the distribution function of the response (here, lifetime) from LBRC-C data;
2. The risk minimization problem, under length bias and right censoring, which comprises the inferential engine of statistical learning theory;
3. Estimation of the regression function of the lifetime from LBRC-C data.

4.1 Learning the Distribution Function from LBRC-C Data

The particular problem of interest in this section is to learn the *distribution function* of the survival time, from a limited sample of LBRC-C data. In other words, we would like to estimate the underlying, unknown probability measure, defined on a certain sample space with a particular set of measurable subsets, when a set of i.i.d., and LBRC-C data are given.

In order to understand the LBRC-C setting better and, especially, to see its difference with the case where the sample data are *representative* of the entire population, i.e., where there is no sampling bias and censoring, let us begin with the following simpler case and then move onto the situation of interest.

Let (Ω, Σ, P) be the probability space of interest. Generally, the sample space Ω can be a set of vectors including, e.g., inputs, outputs, and other related information. With no loss of generality, and in order to formalize the problem mathematically, let us assume that the sample space contains only the covariates and the response variable of interest, i.e., $\Omega = \mathcal{X} \times \mathcal{Y}$. Given that Ω is a product space, the natural choice for the σ-algebra is the tensor-product σ-algebra, i.e., $\Sigma = \Sigma_{\mathcal{X}} \otimes \Sigma_{\mathcal{Y}}$, where $\Sigma_{\mathcal{X}}$ and $\Sigma_{\mathcal{Y}}$ are the individual σ-algebras considered on $\mathcal{X}$ and $\mathcal{Y}$, respectively. Similarly, the probability measure P is defined to be the product measure $P = P_{\mathbf{X}} \times P_{Y}$ with $P_{\mathbf{X}}$ and P_{Y} being the measures defined on $(\mathcal{X}, \Sigma_{\mathcal{X}})$ and $(\mathcal{Y}, \Sigma_{\mathcal{Y}})$, respectively.

The problem of estimating $P(A)$, for any measurable set $A \in \Sigma$, when the measurable space (Ω, Σ) as well as a *limited* number of i.i.d. training examples $\mathcal{D} = \{\mathbf{z}_i = (\mathbf{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y} : i = 1, 2, \ldots, n\}$ are given, is a fundamental problem in mathematical statistics. Let $\hat{P}_n$ denote the estimator of P based on $\mathcal{D}$. In general, $\hat{P}_n$ is called the empirical measure (see Definition 9). Now, having the fundamental problem of mathematical statistics defined, we are ready to take the first step towards the definition of the central problem of interest in the current section, i.e., the problem of *learning (estimating)* the *distribution function* from LBRC-C data. This is the first step, because as mentioned earlier, we first define the problem in the unbiased and complete data setting and then move onto the specific case of LBRC-C data.

The problem of estimating the distribution function might be defined as a special case of the more general fundamental problem of mathematical statistics defined above: Given the measurable space (Ω, Σ), estimate $P(A)$, for any measurable subset $A \in \Sigma_{\preccurlyeq} \subset \Sigma$, where A is of the particular form $A = A_{\mathcal{X}} \times A_{\mathcal{Y}}$ with

$$\begin{aligned} A_{\mathcal{X}} &= \{a \in \mathcal{X} : \mathbf{X}(a) \preccurlyeq \mathbf{x}\}, \quad \mathbf{x} \in \mathbb{R}^d, \\ A_{\mathcal{Y}} &= \{b \in \mathcal{Y} : Y(b) \leq y\}, \qquad y \in \mathbb{R}, \end{aligned} \tag{4.1}$$

where $\preccurlyeq$ is an element-wise inequality. In fact, the σ-algebra Σ has been replaced with one of its proper subsets, i.e., $\Sigma_{\preccurlyeq}$. Note that this new set of measurable subsets does not necessarily need to compose a proper σ-algebra. This is the weak mode of estimation explained in Chapter 3. Moreover, with this particular form assumed, the probability measure P reduces to the CDF.

In the ordinary case of unbiased and complete data, the uniform consistency of the empirical distribution function, as an empirical measure for estimating the underlying distribution, is the direct result of the Glivenko-Cantelli theorem for random vectors (Theorem 4). Notice that, under this scenario, the learning procedure is *fully* supervised as there is no bias and censoring involved. That is, all the values of the response variable Y in the training data are

given.

Now, as it was promised earlier, let us consider the problem of learning the or estimating the distribution function in the specific situation of LBRC-C data. The length bias, formally, can be interpreted as a second probability measure, say $\tilde{P}$, that is defined on the same measurable space $(\Omega, \Sigma_{\preccurlyeq})$. As it was explained before, this is the distribution according to which the realizations of the sample space Ω are observed. On the other hand, censoring does not change the underlying stochastic structure of the experiment.

Since in most practical cases, the target distribution to be estimated is the *unbiased* distribution P, the learning problem in the context of LBRC-C data is almost the same as in the ordinary case except for a slight modification: Let the measurable space $(\Omega, \Sigma_{\preccurlyeq})$ is given. Also, a set of i.i.d. samples $\mathbf{Z}_i$, $i = 1, 2, \ldots, n$, is available. However, this time the data are distributed according to the length-biased distribution, i.e., $\mathbf{Z}_i \sim \tilde{P}$. Additionally, the observations are subject to right censoring. Then, the learning problem is to estimate the *unbiased* distribution function P using the given length-biased and right-censored training data $\mathcal{D}$.

Finally, as the probability measure of interest is the distribution function, i.e., the CDF, we will replace P with the more conventional notation F, which denotes a generic CDF.

Statistical Learning Interpretation of the Problem

In Chapter 3, we have emphasized that, in the framework of statistical learning theory, the general learning problem is a risk minimization problem. Although it might seem *not* obvious, estimating the distribution function can also be formalized as a risk minimization problem. As a matter of fact, this problem is closely related to the *density estimation*. This relation has been thoroughly discussed by Vapnik in different places, including Vapnik and Stepanyuk [1978] and Vapnik [1998]. Particularly, it has been illustrated that how estimating the distribution function or the related density estimation might be expressed as either a risk minimization problem or the so-called problem of *interpreting the results of indirect measurements*. Besides, Vapnik discusses the parametric framework of density estimation, which he calls the Fisher-Wald's setting. In the latter case, the estimation procedure is based on the frequently used MLE, which can be easily expressed in terms of risk minimization. (For details, see section 1.5 of Vapnik [1998].) Our perspective of the problem, however, is based on a *non-parametric* approach.

4.1.1 Preliminaries

We shall formally consider the following two types of data: First, when the sample data contain *no* covariates, i.e., LBRC data, and second, when each subject in the sample data is associated with a vector of covariates, giving rise to LBRC-C data. In both scenarios, the learning procedure is *weakly* supervised as information is available only *partially* due to the right censoring and length bias.

In the context of LBRC-C data, the problem of survival estimation in the first setting above, i.e., when there is no covariates involved, has been discussed thoroughly by Asgharian et al. [2002], Asgharian and Wolfson [2005]. In particular, the asymptotic properties of the estimated, the nonparametric maximum likelihood estimation (NPMLE), survivor function has been completely established. In contrast, learning the survival distribution function in presence of covariates has not been studied yet. So, the core objective of the rest of this section is to investigate the problem of estimating the survival distribution function with LBRC data, including covariates, i.e., LBRC-C data.

First of all, we introduce, and briefly discuss, the background conditions and assumptions we are going to consider throughout our investigation. The notation and basic concepts explained in the following paragraphs are quite general and are assumed to hold in both the LBRC and LBRC-C settings.

Suppose that n subjects from the prevalent population are recruited at the beginning of a followup study. Subjects are collected independently and are planned to be followed up for a certain period of time when the study ends. Of the entire sample, n_f subjects fail during the study, i.e., between the recruitment and end of the study. Another portion of the recruited sample are right censored during the study; these are the individuals who are lost to follow up. Finally, those who survive up to the end of the study are right censored once the study period is over. Assume n_c is the total number of censored subjects. Hence, $n = n_f + n_c$. We will assume that censoring of the residual lifetimes are carried out randomly, i.e., n_f and n_c are both random quantities.

To each observed subject i is attached a quadruple $(\overset{*}{\mathbf{X}}_i, \tilde{A}_i, \tilde{R}_i \wedge C_i, \delta_i)$, where $\overset{*}{\mathbf{X}}_i$ is a vector of covariates, $\tilde{A}_i$, and $\tilde{R}_i$ are the current and residual lifetimes, C_i is the residual censoring time, and δ_i is the failure indicator, defined in subsection 3.3.5.

Clearly, the survival time $\tilde{Y}_i$ is the sum of the current and residual lifetimes, i.e., $\tilde{Y}_i = \tilde{A}_i + \tilde{R}_i$, for each subject i. Similarly, the (overall) censoring time C'_i is the sum of the current lifetime and the residual censoring time, i.e., $C'_i = \tilde{A}_i + C_i$. Another key assumption we assume to be satisfied is the *stationarity* assumption. That is, the *incident rate* over time is unchanged

and, consequently, a stationary Poisson process can reasonably describe the incidence of the initiating event E_1 or onset. Further, we assume that the residual censoring time C is independent of $(\tilde{A}, \tilde{R})$. Note that this is a reasonable assumption in many applications such as settings with type I censoring. On the other hand, the survival time and the overall censoring time are correlated, because one may easily see that, under the stationarity assumption, we have

$$\mathrm{Cov}(\tilde{Y}, C') = \mathrm{Var}(\tilde{A}) \left[1 + \mathrm{corr}(\tilde{A}, \tilde{R}) \sqrt{\frac{\mathrm{Var}\, \tilde{R}}{\mathrm{Var}\, \tilde{A}}} \right].$$

The positive correlation between $\tilde{Y}$ and C' is a result of the stationarity assumption. The reason is that, under stationarity, we have that $\tilde{A}_{|\tilde{Y}} \sim \mathrm{Unif}(0, \tilde{Y})$, which in turn implies that $\mathrm{Var}(\tilde{A}) = \mathrm{Var}(\tilde{Y} - \tilde{A}) = \mathrm{Var}(\tilde{R})$. Therefore, the censoring under consideration is *informative*. We remind the reader that censoring is called *non-informative* or *random* if it is independent of the survival time, which is not the case here.

Before discussing the problem of estimating the distribution function based on LBRC-C data, note that we consider two separate cases:

- *Case One: No Censoring.* Here, the sampled data are assumed to be purely length biased, however, information on the survival of all sampled subjects is fully provided, That is, $n_c = 0$.
- *Case Two: With Censoring.* In this case, in addition to the length bias caused by the sampling procedure, subjects might be right censored. Hence, the exact survival value, for a subset of the sampled subjects, is known only partially, i.e., solely a lower bound for the potential survival time is available. Therefore, $0 < n_c \leq n$.

As shown later, these two cases have essential differences and should be treated separately. In fact, learning the distribution, when data are censored, is harder compared to the case of pure length-biased data.

4.1.2 Case One: No Censoring

Without censoring, $\delta_i = 1$ and $\tilde{R}_i \wedge C_i = \tilde{R}_i$, for all $i = 1, 2, \ldots, n$. That is, one can assume that to any subject i is associated $(\overset{*}{\mathbf{X}}_i, \tilde{A}_i, \tilde{R}_i)$, which is an equivalent but shorter notation than the one introduced earlier. It is easy to see that estimating the length biased distribution function, i.e., $F_{\overset{*}{\mathbf{Z}}|\overline{T}}(\mathbf{z}) = P(\overset{*}{\mathbf{Z}} \preccurlyeq \mathbf{z})$, where $\mathbf{z}$ is a realized value of $\overset{*}{\mathbf{Z}} = (\overset{*}{\mathbf{X}}, \tilde{Y})$ with $\tilde{Y} = \tilde{A} + \tilde{R}$ being the overall lifetime, is rather straightforward. This is because the sample data are distributed according to the same length-biased distribution, which is the

target of the estimation in this case. More precisely, we have that

$$\begin{aligned} F_{\overset{*}{\mathbf{Z}}|\overline{T}}(\mathbf{z}) &= P(\overset{*}{\mathbf{Z}} \preccurlyeq \mathbf{z}) = P(\overset{*}{\mathbf{X}} \preccurlyeq \mathbf{x}, \tilde{Y} \leq y) \\ &= \int\limits_{(\mathbf{u},v)\preccurlyeq(\mathbf{x},y)} \mathrm{d}F_{\overset{*}{\mathbf{X}},\tilde{Y}|\overline{T}}(\mathbf{u},v) = \int\limits_{\mathbf{w}\preccurlyeq\mathbf{z}} \mathrm{d}F_{\overset{*}{\mathbf{Z}}|\overline{T}}(\mathbf{w}) \end{aligned}$$

which depends completely on the biased distributions. Therefore, the multidimensional empirical distribution

$$\begin{aligned} \hat{F}_{\overset{*}{\mathbf{Z}};\overset{*}{n}}(\mathbf{z}) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(\mathbf{z}_i \preccurlyeq \mathbf{z}) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(\mathbf{x}_i \preccurlyeq \mathbf{x}, a_i + r_i \leq y) \end{aligned} \tag{4.2}$$

provides a natural estimator for the desired (length biased) distribution function. The subscript $\overset{*}{\mathbf{Z}};\overset{*}{n}$ highlights the fact that this is an estimator of the *biased* variable $\overset{*}{\mathbf{Z}}$, built upon a *biased* sample of size n, i.e., $\overset{*}{n}$. Furthermore, the Glivenko-Cantelli theorem guarantees that

$$\|\hat{F}_{\overset{*}{\mathbf{Z}};\overset{*}{n}} - F_{\overset{*}{\mathbf{Z}}|\overline{T}}\|_{\infty} \xrightarrow{a.s.} 0, \qquad \text{as } n \to \infty. \tag{4.3}$$

Now, let us consider the estimation of the *unbiased* distribution $F_{\mathbf{Z}}(\mathbf{z})$ under length bias. In fact, what we are interested in is $P(\mathbf{Z} \preccurlyeq \mathbf{z})$, while the data are sampled through a length biased sampling procedure. In contrast to the estimation of $F_{\overset{*}{\mathbf{Z}}|\overline{T}}$, the sample data cannot be naively used in order to construct the empirical measure. Note that, in presence of covariates, the following two identities hold:

$$f_{\tilde{Y}|\mathbf{X},\overline{T}}(y|\mathbf{X}=\mathbf{x}) = \frac{y f_{Y|\mathbf{X}}(y|\mathbf{X}=\mathbf{x})}{\mu(\mathbf{x})}, \tag{4.4}$$

and[1]

$$f_{\overset{*}{\mathbf{X}}|\overline{T}}(\mathbf{x}) = \frac{\mu(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x})}{\mu}, \tag{4.5}$$

[1]There is a subtle point in the notation used in equation (4.4) that might seem confusing: Although we know that the length bias in the response variable leads to the covariates being biased too, the conditional distributions on both sides of the equation are conditioned on $\mathbf{X}$, rather than $\overset{*}{\mathbf{X}}$. The reason is that when the covariate is given, its distribution does not play a role anymore. Hence, the original distribution of the covariates should be used.

where $\mu(\mathbf{x}) := \mathbb{E}_{Y|\mathbf{X}}(y|\mathbf{X}=\mathbf{x})$, and $\mu := \mathbb{E}_{\mathbf{X}}[\mu(\mathbf{x})]$, both in the incident population. The above equations imply that

$$\begin{aligned} f^{*}_{\mathbf{Z}|\overline{T}}(\mathbf{z}) &= f^{*}_{\mathbf{X},\tilde{Y}|\overline{T}}(\mathbf{x},y) \\ &= \frac{y}{\mu} f_{\mathbf{X},Y}(\mathbf{x},y) = \frac{y}{\mu} f_{\mathbf{Z}}(\mathbf{z}). \end{aligned} \tag{4.6}$$

Now, consider the target distribution function to be estimated, i.e., $F_{\mathbf{Z}}(\mathbf{z})$:

$$\begin{aligned} F_{\mathbf{Z}}(\mathbf{z}) &= P(\mathbf{Z} \preccurlyeq \mathbf{z}) = P(\mathbf{X} \preccurlyeq \mathbf{x}, Y \leq y) \\ &= \int\limits_{(\mathbf{u},v)\preccurlyeq(\mathbf{x},y)} \mathrm{d}F_{\mathbf{X},Y}(\mathbf{u},v) = \int\limits_{\mathbf{w}\preccurlyeq\mathbf{z}} \mathrm{d}F_{\mathbf{Z}}(\mathbf{w}), \end{aligned}$$

which, after plugging equation (4.6), yields

$$\begin{aligned} F_{\mathbf{Z}}(\mathbf{z}) &= \int\limits_{(\mathbf{u},v)\preccurlyeq(\mathbf{x},y)} f_{\mathbf{X},Y}(\mathbf{u},v)\,\mathrm{d}(\mathbf{u},v) \\ &= \mu \int\limits_{(\mathbf{u},v)\preccurlyeq(\mathbf{x},y)} v^{-1}\,\mathrm{d}F^{*}_{\mathbf{X},\tilde{Y}|\overline{T}}(\mathbf{u},v) = \mu \int\limits_{\mathbf{w}\preccurlyeq\mathbf{z}} v^{-1}\,\mathrm{d}F^{*}_{\mathbf{Z}|\overline{T}}(\mathbf{w}). \end{aligned} \tag{4.7}$$

As one can see, the last integral above is taken with respect to the biased joint distribution function $F^{*}_{\mathbf{Z}|\overline{T}}$ and can be easily estimated from the sample data. But we still need to estimate the overall mean response μ. In fact, there are two ways to estimate μ from the biased sample. First, using equation (4.5) one can easily derive the following:

$$\mu = \left[\int \frac{1}{\mu(\mathbf{u})} f^{*}_{\mathbf{X}|\overline{T}}(\mathbf{u})\,\mathrm{d}\mathbf{u}\right]^{-1} = \left\{\mathbb{E}^{*}_{\mathbf{X}|\overline{T}}\left[\frac{1}{\mu(\mathbf{x})}\right]\right\}^{-1}, \tag{4.8}$$

which is also based on the biased distribution of the covariates. Replacing μ in equation (4.7) with what equation (4.8) provides, we obtain

$$F_{\mathbf{Z}}(\mathbf{z}) = \frac{\mathbb{E}^{*}_{\mathbf{Z}|\overline{T}}(y^{-1} \mid \overset{*}{\mathbf{Z}} \preccurlyeq \mathbf{z})}{\mathbb{E}^{*}_{\mathbf{X}|\overline{T}}\left[\frac{1}{\mu(\mathbf{x})}\right]}. \tag{4.9}$$

Apparently, the last equation implies that one is able to rewrite the unbiased distribution function, solely, based on the biased joint distribution of the covariates and the response and the biased distribution of the covariate. In other words, $F_{\mathbf{Z}}$ can be estimated using the given LBRC-C sample. However, there is a subtle point here that needs further attention. Recall that $\mu(\mathbf{x}) = \mathbb{E}_{Y|\mathbf{X}}(y|\mathbf{X}=\mathbf{x})$. That is, we still need the *unbiased* distribution of the response

given the covariate to be able to compute this conditional expectation. This information is not readily available, which means $\mu(\mathbf{x})$ itself should be estimated from data. In general, this is not a straightforward estimation, especially, if the covariate is continuous.

Fortunately, this is not the only way one can estimate μ. Note that equation (4.6), also, can be used for this purpose. It is not difficult to see that (4.6) implies

$$\mu = \left[\int v^{-1}\, \mathrm{d}F^{*}_{\mathbf{Z}|\overline{T}}(\mathbf{u}, v)\right]^{-1} = \left[\mathbb{E}^{*}_{\mathbf{Z}|\overline{T}}(y^{-1})\right]^{-1}, \tag{4.10}$$

which can be, directly, estimated from the biased data. Now, equation (4.10) together with (4.7) provide one with the following expression:

$$F_{\mathbf{Z}}(\mathbf{z}) = \frac{\mathbb{E}^{*}_{\mathbf{Z}|\overline{T}}(y^{-1} \mid \overset{*}{\mathbf{Z}} \preccurlyeq \mathbf{z})}{\mathbb{E}^{*}_{\mathbf{Z}|\overline{T}}(y^{-1})}.$$

The last equation, however, gives rise to the following natural empirical distribution:

$$\hat{F}_{\mathbf{Z};\overset{*}{n}}(\mathbf{z}) = \frac{n}{n_{\mathbf{z}}}\left[\sum_{i=1}^{n} y_i^{-1}\, \mathbb{1}(\mathbf{z}_i \preccurlyeq \mathbf{z})\right]\left(\sum_{i=1}^{n} y_i^{-1}\right)^{-1} \tag{4.11}$$

where $n_{\mathbf{z}} = \sum_{i=1}^{n} \mathbb{1}(\mathbf{z}_i \preccurlyeq \mathbf{z})$. The subscript $\mathbf{Z}, \overset{*}{n}$ is to emphasize that this is an estimator of the *unbiased* variable $\mathbf{Z}$, based on a sample dataset consisting of n, *biased* subjects. Equation (4.11) does not require the additional estimation of $\mu(\mathbf{x})$, in contrast with the estimator one could have driven based on equation (4.9).

Naturally, after constructing an estimator, one would like to check its consistency. Here, we skip the technical details of a proper proof of consistency and shortly provide a sketch of the proof and the general idea behind it.

The almost sure consistency of the estimator (4.11) is implied from the fact that $F_{\mathbf{Z},\overset{*}{n}}$, obtained above, is a *continuous* transformation of the sequence $F_{\overset{*}{\mathbf{Z}},\overset{*}{n}}$, defined earlier by equation (4.2), together with the Glivenko-Cantelli result stated in equation (4.3).

4.1.3 Case Two: With Censoring

As stated before, a tuple $(\overset{*}{\mathbf{X}}_i, \tilde{A}_i, \tilde{R}_i \wedge C_i, \delta_i)$ is associated to every observed subject i. Under the stationarity assumption, the joint density of the current and residual lifetimes, condi-

tioned on the covariates and the left truncation, is given by

$$f_{\tilde{A},\tilde{R}|\mathbf{X},\overline{T}}(a,r \mid \mathbf{X}=\mathbf{x}) = \begin{cases} \dfrac{f_{Y|\mathbf{X}}(a+r \mid \mathbf{X}=\mathbf{x})}{\mathbb{E}(Y|\mathbf{X}=\mathbf{x})}, & a,r>0, \\ 0, & \text{otherwise}, \end{cases} \tag{4.12}$$

where $\overline{T}$ denotes the left truncation [Vardi, 1989, Feller, 1971].

The final objective is to learn the CDF of the unbiased joint distribution, i.e., $F_{\mathbf{Z}}(\mathbf{z})$, for any $\mathbf{z} \in \mathbb{R}^{d+1}$. Since the relation between the unbiased and length-biased distribution of the lifetime is known, we can achieve the goal of learning the unbiased distribution through learning the length-biased one first and then transform it to the desired unbiased distribution; that is, first, learning $F^{*}_{\mathbf{Z}|\overline{T}}$, and further use it to find $F_{\mathbf{Z}}$.

In the following paragraphs, we first derive the nonparametric maximum likelihood function, and then, following a similar procedure as Asgharian and Wolfson [2005], will try to estimate the length-biased distribution function. Note that what makes the situation different here, from the one considered in Asgharian and Wolfson [2005], is the presence of covariates in the present setting.

Consider the sub-sample of the *failed* subjects and define the following:

$$\tilde{F}_1(\mathbf{z}) := F^{*}_{\mathbf{Z}|\delta,\overline{T}}(\mathbf{z} \mid \delta=1) = \mathbb{P}(\overset{*}{\mathbf{X}} \preccurlyeq \mathbf{x}, \tilde{A}+\tilde{R} \leq y \mid \delta=1).$$

Let $\tilde{f}_1$ be the corresponding density. Then,

$$\begin{aligned} \tilde{f}_1(\mathbf{z}) &= f^{*}_{\mathbf{Z}|\delta,\overline{T}}(\overset{*}{\mathbf{X}}=\mathbf{x}, \underbrace{\tilde{A}+\tilde{R}}_{\tilde{Y}} = \underbrace{a+r}_{y} \mid \delta=1) \\ &= \frac{f^{*}_{\mathbf{Z},\delta|\overline{T}}(\overset{*}{\mathbf{X}}=\mathbf{x}, \tilde{A}+\tilde{R}=a+r, \delta=1)}{\mathbb{P}_{\delta|\overline{T}}(\delta=1)}. \end{aligned}$$

Let $\mathbb{P}_{\delta|\overline{T}}(\delta=1) = \mathbb{P}(\delta=1 \mid \overline{T}) = p$. Hence, by the law of total probability we obtain

$$\begin{aligned} \tilde{f}_1(\mathbf{z}) &= \frac{1}{p} f^{*}_{\mathbf{Z},C|\overline{T}}(\overset{*}{\mathbf{X}}=\mathbf{x}, \tilde{A}+\tilde{R}=a+r, \underbrace{C>\tilde{R}}_{\delta=1}) \\ &= \frac{1}{p}\int_0^y f^{*}_{\mathbf{Z},C|\tilde{R},\overline{T}}(\overset{*}{\mathbf{X}}=\mathbf{x}, \tilde{A}+\tilde{R}=a+r, C>\tilde{R} \mid \tilde{R}=r) f_{\tilde{R}|\overline{T}}(r)\,\mathrm{d}r \\ &= \frac{1}{p}\int_0^y f^{*}_{\mathbf{X},\tilde{A},C|\tilde{R},\overline{T}}(\overset{*}{\mathbf{X}}=\mathbf{x}, \tilde{A}=a, C>r \mid \tilde{R}=r) f_{\tilde{R}|\overline{T}}(r)\,\mathrm{d}r. \end{aligned}$$

The residual censoring was assumed to be independent from the current lifetime as well as the covariates, which implies that

$$\begin{aligned}\tilde{f}_1(\mathbf{z}) &= \frac{1}{p}\int_0^y f_{\overset{*}{\mathbf{X}},\tilde{A}|\tilde{R},\overline{T}}(\overset{*}{\mathbf{X}}=\mathbf{x},\tilde{A}=a\mid\tilde{R}=r)\,\mathbb{P}_{C|\overline{T}}(C>r)f_{\tilde{R}|\overline{T}}(r)\,\mathrm{d}r\\ &= \frac{1}{p}\int_0^y f_{\overset{*}{\mathbf{X}},\tilde{A},\tilde{R}|\overline{T}}(\overset{*}{\mathbf{X}}=\mathbf{x},\tilde{A}=a,\tilde{R}=r)\,S_{C}(r)\,\mathrm{d}r\\ &= \frac{1}{p}\int_0^y f_{\tilde{A},\tilde{R}|\mathbf{X},\overline{T}}(\tilde{A}=a,\tilde{R}=r\mid\mathbf{X}=\mathbf{x})\,f_{\overset{*}{\mathbf{X}}|\overline{T}}(\mathbf{x})\,S_{C}(r)\,\mathrm{d}r,\end{aligned}$$

where $S_{C}(r)=1-F_{C}(r)$ is the survival function of the residual censoring time. Now, plugging equations (4.12) and (4.5) yields

$$\begin{aligned}\tilde{f}_1(\mathbf{z}) &= f_{\overset{*}{\mathbf{Z}}|\delta,\overline{T}}(\mathbf{z}\mid\delta=1)\\ &= \frac{1}{p}\int_0^y \frac{f_{Y|\mathbf{X}}(a+r\mid\mathbf{X}=\mathbf{x})f_{\mathbf{X}}(\mathbf{x})}{\mu}\,S_{C}(r)\,\mathrm{d}r\\ &= \frac{f_{\mathbf{Z}}(\mathbf{z})}{p\,\mu}\int_0^y S_{C}(r)\,\mathrm{d}r,\end{aligned}\tag{4.13}$$

which shows the relation between the unbiased joint density $f_{\mathbf{X},Y}$ with the biased joint density of the *failed* sampled subjects $\tilde{f}_1$. For ease, denote $\zeta(y):=\int_0^y S_{C}(r)\,\mathrm{d}r$. Then, from the last equation one can derive the corresponding distribution function:

$$\begin{aligned}F_{\mathbf{Z}}(\mathbf{z}) &= \int\limits_{\mathbf{w}\preccurlyeq\mathbf{z}} f_{\mathbf{Z}}(\mathbf{w})\,\mathrm{d}\mathbf{w}\\ &= p\,\mu\int\limits_{\mathbf{w}\preccurlyeq\mathbf{z}} [\zeta(v)]^{-1}\,\mathrm{d}\tilde{F}_1(\mathbf{w}).\end{aligned}\tag{4.14}$$

Now, in order for one to be able to derive the empirical distribution function based on equation (4.14), μ and ζ, also, need to be expressed based on the biased joint distribution since the available data are biased. Using equation (4.13), one can also verify that

$$\mu = p\left\{\int_{\mathbf{w}}[\zeta(v)]^{-1}\,\mathrm{d}\tilde{F}_1(\mathbf{w})\right\}^{-1},\tag{4.15}$$

plugging which in equation (4.14) yields that

$$\begin{aligned} F_{\mathbf{Z}}(\mathbf{z}) &= p^2 \left\{ \int_{\mathbf{w}} [\zeta(v)]^{-1} \, \mathrm{d}\tilde{F}_{1}(\mathbf{w}) \right\}^{-1} \left\{ \int_{\mathbf{w} \preccurlyeq \mathbf{z}} [\zeta(v)]^{-1} \, \mathrm{d}\tilde{F}_{1}(\mathbf{w}) \right\} \\ &= p^2 \, \frac{\mathbb{E}^{*}_{\mathbf{Z}|\delta,\overline{T}} \left\{ \left[\zeta(\tilde{Y}) \right]^{-1} \mid \delta = 1, \overset{*}{\mathbf{Z}} \preccurlyeq \mathbf{z} \right\}}{\mathbb{E}^{*}_{\mathbf{Z}|\delta,\overline{T}} \left\{ \left[\zeta(\tilde{Y}) \right]^{-1} \mid \delta = 1 \right\}}. \end{aligned} \tag{4.16}$$

The last equation, literally, means that the unbiased joint distribution function $F_{\mathbf{Z}}$ might be expressed as a transformation of the biased joint distribution $\tilde{F}_{1}$ and the survival function of the residual censoring, i.e., S_{C}. More precisely, we have that

$$F_{\mathbf{Z}} = F_{\mathbf{X},Y} = \Gamma(S_{C}, \tilde{F}_{1}),$$

where

$$\begin{aligned} \Gamma\left[\zeta(y), G(\mathbf{z})\right] = p^2 &\left\{ \int_{\mathbf{w}} \left[\int_{0}^{v} S_{C}(r) \, \mathrm{d}r \right]^{-1} \mathrm{d}G(\mathbf{w}) \right\}^{-1} \\ &\times \left\{ \int_{\mathbf{w} \preccurlyeq \mathbf{z}} \left[\int_{0}^{v} S_{C}(r) \, \mathrm{d}r \right]^{-1} \mathrm{d}G(\mathbf{w}) \right\}, \end{aligned} \tag{4.17}$$

defines the transformation Γ, for any functions $S : \mathbb{R} \to [0,1]$ and $G : \mathbb{R}^{d+1} \to [0,1]$. Note that Γ is a continuous transformation, in both arguments. Now, based on equation (4.16), the empirical distribution function will be of the following form:

$$\hat{F}_{\mathbf{Z},n}^{*}(\mathbf{z}) = \hat{p}_{n}^{2} \, \frac{n_{f}}{n_{\mathbf{z},f}} \sum_{i=1}^{n} \left\{ \left[\int_{0}^{y_i} \hat{S}_{C}(r) \, \mathrm{d}r \right]^{-1} \delta_{i} \, \mathbb{1}(\mathbf{z}_{i} \preccurlyeq \mathbf{z}) \right\} \left\{ \sum_{i=1}^{n} \delta_{i} \left[\int_{0}^{y_i} \hat{S}_{C}(r) \, \mathrm{d}r \right]^{-1} \right\}^{-1}, \tag{4.18}$$

where $\hat{p}_{n}$ is the estimate of p, i.e., $\mathbb{P}(\delta = 1 \mid \overline{T})$, $n_{f} = \sum_{i=1}^{n} \delta_{i}$, and $n_{\mathbf{z},f} = \sum_{i=1}^{n} \delta_{i} \, \mathbb{1}(\mathbf{z}_{i} \preccurlyeq \mathbf{z})$.

As one can see in equation (4.18), the survival function of the residual censoring times $S_{C}(r)$, also, needs to be estimated from data. The good news is that the well-known *Kaplan-Meier* (aka *product limit*) estimator might be applied to this end. Traditionally, the Kaplan-Meier estimator is used to, non-parametrically, estimate the survival function of the lifetime, and specifically, it can take care of right censoring [Kaplan and Meier, 1958]. The Kaplan-Meier estimator has been discussed in the literature extensively and, hence, we skip the details. However, we briefly introduce how to apply it for estimating S_{C}.

First, notice that the aforementioned traditional context is not exactly what we are looking for: In connection with equation (4.18), what we are interested in is the survival function of the *residual censoring time* and not the survival function of the *lifetime*. Interestingly, besides its conventional context, the Kaplan-Meier estimator can be used in a broader sense as explained further: The event of interest does not necessarily require to be the failure time. In fact, in a more general setting, we assume that there exist two hypothetical "terminating" events, say E_2 and E_2', any of which can occur first. The key point here is that, for any subject, only the quantity $E_2 \wedge E_2'$ could be observed, i.e., the event which occurs first. The Kaplan-Meier estimator might be equally applied to estimate the survival function of either of E_2 or E_2', depending on the desired objective.

Returning to our problem, we are interested in the survival function of the residual censoring time, nonetheless, for a portion of the subjects the exact censoring time is not available due to failure happen first. In other words, the right censoring time itself is subject to being right censored, due to a sooner failure. Therefore, the Kaplan-Meier estimator of S_C can be written as follows:

$$\hat{S}_C(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right),$$

where t_i's are the moments at which at least one of the events E_2, E_2' occurs, d_i is the number of the specific event of interest, i.e., right-censored subjects at t_i, and n_i is the total number of subjects who have not experienced neither E_2 nor E_2' at t_i. Figure 4.1 provides a toy example of the Kaplan-Meier curve.

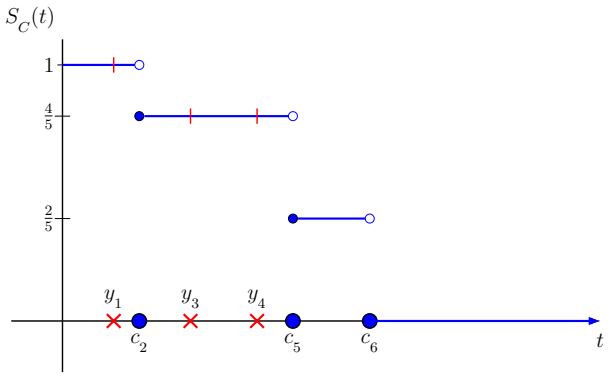

Figure 4.1 **Kaplan-Meier Estimator of the Survival of the Total Censoring Time.** Here, the step function in blue illustrates the empirical estimate of the survival function of the residual censoring time, i.e., S_C, for example data consisting of 3 censored subjects at times c_2, c_5 and c_6 (red crosses), and 3 failed subjects at times y_1, y_3, and y_4 (red crosses). Vertical red tick-marks show occurrence of the alternative event, i.e., failure.

This completes the estimation of the distribution function $F_{\mathbf{Z}}$ based on LBRC-C sample

data. An interesting fact about estimator (4.18) is that it was derived purely based on the sub-sample of the failed subjects. In practice, this may lead to efficiency loss, especially, if data are heavily right censored. Nevertheless, when the censored portion of the sample is negligible, equation (4.18) constitutes a reasonably efficient estimator of the distribution function. Next, the consistency of estimator will be discussed.

Theorem 5 (Uniform Almost-Sure Consistency of the Empirical Distribution Function with LBRC-C Data)**.** *Let $F_{\mathbf{Z},n}^{*}$ denote the estimator defined by equation (4.18). Then, $F_{\mathbf{Z},n}^{*}$, calculated from LBRC-C sample data, constitutes a uniform, almost-surely, consistent estimator of the unbiased distribution function $F_{\mathbf{Z}}$, given by equation (4.14). That is,*

$$\|F_{\mathbf{Z};n}^{*} - F_{\mathbf{Z}}\|_{\infty} \xrightarrow{a.s.} 0, \qquad as\ n \to \infty.$$

Proof. First, notice that

$$F_{\mathbf{Z};n}^{*} = \Gamma_{n}(\hat{S}_{C}, \hat{\tilde{F}}_{1}),$$

where Γ_{n} is the empirical version of the continuous transformation Γ defined by equation (4.17). Denote

$$\lambda_{n}(\mathbf{z}) := \frac{1}{n_{\mathbf{z},1}} \left\{ \sum_{i=1}^{n} \left[\int_{0}^{y_i} \hat{S}_{C}(r)\, \mathrm{d}r \right]^{-1} \delta_{i}\, \mathbb{1}(\mathbf{z}_{i} \preccurlyeq \mathbf{z}) \right\}$$

$$\lambda'_{n}(\mathbf{z}) := \frac{1}{n_{1}} \left\{ \sum_{i=1}^{n} \delta_{i} \left[\int_{0}^{y_i} \hat{S}_{C}(r)\, \mathrm{d}r \right]^{-1} \right\},$$

with $n_{1} = \sum_{i=1}^{n} \delta_{i}$, and $n_{\mathbf{z},1} = \sum_{i=1}^{n} \delta_{i}\, \mathbb{1}(\mathbf{z}_{i} \preccurlyeq \mathbf{z})$ and

$$\lambda(\mathbf{z}) := \int\limits_{\mathbf{w} \preccurlyeq \mathbf{z}} [\zeta(v)]^{-1}\, \mathrm{d}\tilde{F}_{1}(\mathbf{w})$$

$$\lambda'(\mathbf{z}) := \left\{ \int_{\mathbf{w}} [\zeta(v)]^{-1}\, \mathrm{d}\tilde{F}_{1}(\mathbf{w}) \right\}^{-1}.$$

The Kaplan-Meier estimator $\hat{S}_{C}$ is uniformly, almost surely, consistent, i.e., $\|\hat{S}_{C} - S_{C}\|_{\infty} \xrightarrow{a.s.} 0$, as n tends to infinity. Also, note that $\hat{\tilde{F}}_{1;n}^{*}$ is simply the empirical distribution and, hence, converges to the actual distribution function, i.e., $\|\hat{\tilde{F}}_{1;n}^{*} - \tilde{F}_{1}\|_{\infty} \xrightarrow{a.s.} 0$. Consequently, as

$n \to \infty$, we have that

$$\begin{aligned} \lambda_{n}(\mathbf{z}) &\xrightarrow{a.s.} \lambda(\mathbf{z}), \\ \lambda'_{n}(\mathbf{z}) &\xrightarrow{a.s.} \lambda'(\mathbf{z}). \end{aligned} \tag{4.19}$$

On the other hand, it is easy to see that

$$\begin{aligned} \|F_{\mathbf{Z};n}^{*} - F_{\mathbf{Z}}\|_{\infty} &= \sup_{\mathbf{z}} \left| \Gamma_{n}(\hat{S}_{C}, \hat{\tilde{F}}_{1}) - \Gamma(S_{C}, \tilde{F}_{1}) \right| \\ &= \sup_{\mathbf{z}} \left| \frac{\lambda_{n}(\mathbf{z})}{\lambda'_{n}(\mathbf{z})} - \frac{\lambda(\mathbf{z})}{\lambda'(\mathbf{z})} \right| \\ &\leq \frac{1}{\lambda'(\mathbf{z})} \sup_{\mathbf{z}} |\lambda_{n}(\mathbf{z}) - \lambda(\mathbf{z})| + \frac{\lambda(\mathbf{z})}{\lambda'_{n}(\mathbf{z})\lambda'(\mathbf{z})} \sup_{\mathbf{z}} |\lambda'_{n}(\mathbf{z}) - \lambda'(\mathbf{z})|, \end{aligned}$$

Now, as n grows, the last inequality together with equations (4.19) complete the proof. □

4.2 Risk Minimization under LBRC-C Data

In this section of the current chapter, we investigate the problem of minimizing the risk functional when data are LBRC-C. As we have seen in Chapter 3, risk minimization is the essential inferential machinery of statistical learning theory for solving all major supervised problems [Vapnik, 1998]. Recall that in order to find the optimal hypothesis or functional dependence, one desires to minimize the expected risk functional $R(\boldsymbol{\theta})$ over the set of admissible functions $\boldsymbol{\Theta}$. (See The Learning Problem.)

As discussed in subsection 3.1.3, one of the most important goals of statistical learning theory is to provide the necessary and sufficient conditions for reliable learning. Also, it was explained that the reliability is measured by the concept of non-trivial consistency, which in turn was guaranteed by the two-sided uniform convergence. More precisely, it was shown that the two-sided uniform convergence was sufficient for a learning machine to be non-trivially consistent. Nonetheless, the two-sided convergence was not necessary for non-trivial consistency; as a matter of fact, the one-sided uniform convergence was shown to be the sufficient and necessary condition for it. Here, we establish this problem for the case of LBRC-C data and derive the sufficient conditions for consistent learning from this type of data.

In other words, the present section provides the foundations of a reliable learning procedure from data with a specific type of incompleteness we are interested in. As a result, we provide general learning machines that can be reliably utilized to solve the main supervised learning

problems. It is worth mentioning that while we do not, explicitly, target the classification problem here, once the reliability of the risk minimization problem is established, results can be extended to classification easily since the only difference amongst the main supervised learning problems boils down to the choice of a specific loss function.

This section is closely related to the previous one since risk minimization requires learning the underlying probability measure, in the first place. However, in a broader sense, learning density is a special case of the risk minimization problem.

To begin with, we consider the problem in the setting of Case One, discussed in subsection 4.1.2. That is, we first study the properties of risk minimization when the available data are purely length biased and there is no censoring involved. After, we move onto Case Two, considered in subsection 4.1.3, where data are supposed to exhibit both the sampling bias and right censoring. Note that, although we focus, particularly, on length bias and right censoring in this work, the results achieved can be easily extended to *any* type of sampling bias provided that the relation between the biased and unbiased populations is known.

4.2.1 Expected and Empirical Risk Functionals

First, recall that $Q_{\boldsymbol{\theta}}(\mathbf{Z}) = L[Y, h_{\boldsymbol{\theta}}(\mathbf{X})]$, with L being a loss function, $\mathbf{Z} = (\mathbf{X}, Y)$, and $h_{\boldsymbol{\theta}} \in \mathcal{H}_{\boldsymbol{\Theta}}$. The expected risk to be minimized is the following stochastic process:

$$R(\boldsymbol{\theta}) = F_{\mathbf{Z}}\, Q_{\boldsymbol{\theta}} = \mathbb{E}_{\mathbf{Z}}\left[Q_{\boldsymbol{\theta}}(\mathbf{z})\right] = \int Q_{\boldsymbol{\theta}}\, \mathrm{d}F_{\mathbf{Z}}, \tag{4.20}$$

where $F_{\mathbf{Z}}$ is the *unbiased* actual distribution function. In the context of our interest, i.e., where data are LBRC-C, the empirical risk should naturally be defined accordingly. That is,

$$\hat{R}_{\mathbf{Z};n}^{*}(\boldsymbol{\theta}) = \hat{F}_{\mathbf{Z};n}^{*}\, Q_{\boldsymbol{\theta}} = \mathbb{E}_{\mathbf{Z};n}^{*}\left[Q_{\boldsymbol{\theta}}(\mathbf{z})\right] = \int Q_{\boldsymbol{\theta}}\, \mathrm{d}\hat{F}_{\mathbf{Z};n}^{*}. \tag{4.21}$$

Equations (4.20) and (4.21) provide the general form of the expected and empirical risk functionals, respectively. Next, we will separately give the individual risks for the case of pure length bias and simultaneous length bias and right censoring, i.e., Case One and Case Two.

4.2.2 Risk Minimization with Pure Length Bias

When there is no censoring, equation (4.6), which provides the relation between the biased and unbiased joint distribution, together with equation (4.20) imply that the expected risk

can be expressed as follows:

$$\begin{aligned} R(\boldsymbol{\theta}) &= \int Q_{\boldsymbol{\theta}}(\mathbf{u}, v)\ \mathrm{d}F_{\mathbf{X},Y}(\mathbf{u}, v) \\ &= \mu \int v^{-1} Q_{\boldsymbol{\theta}}(\mathbf{u}, v)\, \mathrm{d}F^{*}_{\mathbf{X},\tilde{Y}|\overline{T}}(\mathbf{u}, v) \\ &= \mu\ \mathbb{E}^{*}_{\mathbf{Z}|\overline{T}}\left[\tilde{Y}^{-1}\, Q_{\boldsymbol{\theta}}(\overset{*}{\mathbf{Z}})\right]. \end{aligned}$$

We have already seen that equation (4.10) provides an estimator for the overall mean lifetime μ that can be estimated from the length biased data. Hence, replacing μ with right-hand side of equation (4.10) brings us to the expected risk functional, fully expressed in terms of the length-biased training data:

$$R(\boldsymbol{\theta}) = \left[\mathbb{E}^{*}_{\mathbf{Z}|\overline{T}}(\tilde{Y}^{-1})\right]^{-1} \mathbb{E}^{*}_{\mathbf{Z}|\overline{T}}\left[\tilde{Y}^{-1}\, Q_{\boldsymbol{\theta}}(\overset{*}{\mathbf{Z}})\right].$$

Now, the corresponding empirical risk can be easily derived form the above equation:

$$\hat{R}_{\mathbf{Z};n}^{*}(\boldsymbol{\theta}) = \left(\sum_{i=1}^{n} y_i^{-1}\right)^{-1} \left[\sum_{j=1}^{n} y_j^{-1} Q_{\boldsymbol{\theta}}(\mathbf{z}_j)\right].$$

4.2.3 Risk Minimization with Length Bias and Right Censoring

Now, if there is right censoring involved, then, applying equation (4.13), to (4.20) gives

$$\begin{aligned} R(\boldsymbol{\theta}) &= \int Q_{\boldsymbol{\theta}}(\mathbf{u}, v)\ \mathrm{d}F_{\mathbf{X},Y}(\mathbf{u}, v) \\ &= p\mu \int Q_{\boldsymbol{\theta}}(\mathbf{u}, v)\, [\zeta(v)]^{-1}\, f^{*}_{\mathbf{X},\tilde{Y}|\delta,\overline{T}}(\mathbf{u}, v \mid \delta = 1)\ \mathrm{d}(\mathbf{u}, v) \\ &= p\mu\ \mathbb{E}^{*}_{\mathbf{Z}|\delta,\overline{T}}\left\{[\zeta(\tilde{Y})]^{-1} Q_{\boldsymbol{\theta}}(\overset{*}{\mathbf{Z}}) \mid \delta = 1\right\}. \end{aligned}$$

Similarly as before, the overall mean μ should be replaced with the right-hand side of equation (4.15). So, the expected risk is given as

$$R(\boldsymbol{\theta}) = p^2 \left(\mathbb{E}^{*}_{\mathbf{Z}|\delta,\overline{T}}\left\{\left[\zeta(\tilde{Y})\right]^{-1} \mid \delta = 1\right\}\right)^{-1} \mathbb{E}^{*}_{\mathbf{Z}|\delta,\overline{T}}\left\{[\zeta(\tilde{Y})]^{-1} Q_{\boldsymbol{\theta}}(\overset{*}{\mathbf{Z}}) \mid \delta = 1\right\},$$

where $\zeta(y) := \int_0^y S_C(r)\, \mathrm{d}r$. Clearly, this last equality explicitly expresses the expected risk functional in terms of the distributional information provided by *failed* individuals in the

prevalent cohort, which ultimately leads us to the following empirical risk:

$$\hat{R}_{\mathbf{Z};n}^{*}(\boldsymbol{\theta}) = \hat{p}_n^2 \left\{ \sum_{i=1}^{n} \delta_i \left[\zeta(y_i) \right]^{-1} \right\}^{-1} \sum_{j=1}^{n} \delta_i \left[\zeta(y_j) \right]^{-1} Q_{\boldsymbol{\theta}}(\mathbf{z}_j).$$

Once we have the concrete forms of the expected and empirical risk, it is time to examine how reliable this machinery is for the purpose of solving the learning problem. In the following subsection, we discuss the consistency and related issues in the LBRC-C setting.

4.2.4 Reliability of Learning from LBRC-C Data

The notation used in previous sections allows expressing the conditions needed for consistency in a unified subsection, rather than having to discuss Case One and Case Two, separately. The general form of the two-sided uniform convergence, which is a sufficient condition for non-trivial consistency, requires a learning machine to satisfy

$$\| \hat{R}_{\mathbf{Z};n}^{*}(\boldsymbol{\theta}) - R(\boldsymbol{\theta}) \|_{\infty} \xrightarrow{a.s.} 0, \qquad \text{as } n \to \infty. \tag{4.22}$$

Hence, the eventual objective is to verify whether the above condition holds for the risk functionals extracted in the preceding subsection. To this end, notice that, for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,

$$\begin{aligned} \hat{R}_{\mathbf{Z};n}^{*}(\boldsymbol{\theta}) - R(\boldsymbol{\theta}) &= \hat{F}_{\mathbf{Z};n}^{*} Q_{\boldsymbol{\theta}} - F_{\mathbf{Z}} Q_{\boldsymbol{\theta}} \\ &= \int Q_{\boldsymbol{\theta}}(\mathbf{z}) \, \mathrm{d}\hat{F}_{\mathbf{Z};n}^{*}(\mathbf{z}) - \int Q_{\boldsymbol{\theta}}(\mathbf{z}) \, \mathrm{d}F_{\mathbf{Z}}(\mathbf{z}) \\ &= \int Q_{\boldsymbol{\theta}}(\mathbf{z}) \, \mathrm{d}\left(\hat{F}_{\mathbf{Z};n}^{*} - F_{\mathbf{Z}} \right)(\mathbf{z}) \, . \end{aligned}$$

Now, opening the inner difference up applying the integration by parts we obtain that

$$\hat{R}_{\mathbf{Z};n}^{*}(\boldsymbol{\theta}) - R(\boldsymbol{\theta}) = Q_{\boldsymbol{\theta}}(\mathbf{z}) \left[\left(\hat{F}_{\mathbf{Z};n}^{*} - F_{\mathbf{Z}} \right)(\mathbf{z}) \right] - \int \left[\left(\hat{F}_{\mathbf{Z};n}^{*} - F_{\mathbf{Z}} \right)(\mathbf{z}) \right] \mathrm{d}Q_{\boldsymbol{\theta}}(\mathbf{z}).$$

Note that integration by parts, in situation above, requires the function $Q_{\boldsymbol{\theta}}$ to be of bounded variation as well as the difference function $\hat{F}_{\mathbf{Z};n}^{*} - F_{\mathbf{Z}}$ to be differentiable almost everywhere. In addition, and with no loss of generality, we assume that $\mathbf{z}$ accepts values in a bounded set. Now, one can check that the last equation implies that

$$\begin{aligned} \| \hat{R}_{\mathbf{Z};n}^{*}(\boldsymbol{\theta}) - R(\boldsymbol{\theta}) \|_{\infty} \leq \| \hat{F}_{\mathbf{Z};n}^{*} - F_{\mathbf{Z}} \|_{\infty} \sup_{\mathbf{z}} |Q_{\boldsymbol{\theta}}(\mathbf{z})| \\ + \| \hat{F}_{\mathbf{Z};n}^{*} - F_{\mathbf{Z}} \|_{\infty} \int |\, \mathrm{d}Q_{\boldsymbol{\theta}}(\mathbf{z})|, \end{aligned}$$

which, in turn, gives the sufficient conditions for the two-sided uniform convergence (4.22). It is easy to see that if

$$\|\hat{F}_{\mathbf{Z};n}^{*} - F_{\mathbf{Z}}\|_{\infty} \xrightarrow{a.s.} 0, \quad \text{as } n \to \infty,$$

as well as

$$\sup_{\mathbf{z}} |Q_{\boldsymbol{\theta}}(\mathbf{z})| < \infty, \text{ and } \int |\,\mathrm{d}Q_{\boldsymbol{\theta}}(\mathbf{z})| < \infty, \tag{4.23}$$

then, convergence (4.22) holds. For the specific risk functionals we have introduced here, based on LBRC-C data, it was already shown (in the preceding section) that, for both Case One and Case Two, $\|\hat{F}_{\mathbf{Z};n}^{*} - F_{\mathbf{Z}}\|_{\infty} \xrightarrow{a.s.} 0$, as n increases. Therefore, conditions (4.23) provide the sufficient conditions for non-trivial consistency of the risk minimization procedure based on the empirical risks we have defined in this subsection. The fortunate fact is that the extracted conditions are quite general and, consequently, hold for most of the commonly used loss functions. Verifying conditions (4.23) for particular losses is rather trivial and is left to the reader.

4.3 Regression Analysis under LBRC-C Data

All problems that have been considered so far in this chapter were related to the estimation of the underlying probability measure using empirical data. Particularly, we have explained that learning the (lifetime) distribution function is a less general case of learning the probability measure defined on an arbitrary σ-algebra of measurable sets. In other words, in all of the discussed cases, the target was to discover the distributional structure of the stochastic experiment of interest *completely*. Recall that the objective was to estimate $P(A)$, for any measurable set A of the specific form given in equations (4.1). In the most general case, it led us to replace P with its empirical counterpart $\hat{P}_n$ that could *estimate* P consistently.

Nevertheless, there are situations where the comprehensive distributional characteristics of a stochastic phenomenon are either impossible or hard to learn, or are not of direct interest. In the remainder of this section, we propose a novel method of estimating the regression function when the data are LBRC-C.

Here, we focus on the non-explicit model of regression estimation. Specifically, we show why applying the Nadaraya-Watson estimator, naively, to LBRC-C data leads to invalid estimates. In addition, we will clearly show how one might correct the estimates.

One of the main subcategories of the supervised learning problems in learning theory belongs to problem of regression estimation. Aligned with the main problem of statistical learning, demonstrated in subsection 3.1.1, it is assumed that each statistical unit consists of a

covariate (or input) vector $\mathbf{X} \in \mathbb{R}^d$ and a response (output) $Y \in \mathbb{R}$ that are related by the stochastic functional dependence $F_{Y|\mathbf{X}}$. As mentioned earlier, for the sake of regression analysis, one wants to solve the easier problem of estimating a certain stochastic property of the distribution $F_{Y|\mathbf{X}}$, e.g., one of its central tendencies, rather than estimation of the distribution function, completely.

In this work, we focus on the study of mean regression, which is learning the following function from data:

$$\mathbb{E}_{Y|\mathbf{X}}(Y = y \mid \mathbf{X} = \mathbf{x}) = \int v \; \mathrm{d}F_{Y|\mathbf{X}}(Y = v \mid \mathbf{X} = \mathbf{x}), \qquad \forall \mathbf{x} \in \mathbb{R}^d.$$

We call the left-hand side of the equation above, $\mathbb{E}_{Y|\mathbf{X}}(Y = y \mid \mathbf{X} = \mathbf{x})$, or shortly $\mathbb{E}_{Y|\mathbf{X}}(y \mid \mathbf{x})$, the *regression function* and the regression problem is defined as *estimation of the regression function*. Recall that the available data come from the joint distribution $F_{\mathbf{X},Y}$, however.

Existing approaches to the regression problem, has been divided into different groups in multiple ways. The popular division of the regression analysis models (or methods) into *parametric* and *nonparametric* is, perhaps, the most frequently used classification. Nevertheless, we avoid using this terminology because of its ambiguity [Le Cam and Lo Yang, 1990]; particularly, parametric and non-parametric in the context of regression differ from what Vapnik means, when he described the shortcomings of the classical paradigm of statistical inference. (See Chapter 3.)

Instead, we use the terms *explicit (definite)* and *non-explicit (indefinite)* in the sense explained below: If the regression function is assumed to be a parametric function of the covariates, i.e.,

$$\mathbb{E}_{Y|\mathbf{X}}(Y = y \mid \mathbf{X} = \mathbf{x}) = r_{\boldsymbol{\theta}}(\mathbf{x}), \qquad \boldsymbol{\theta} \in \boldsymbol{\Theta}, \, \forall \mathbf{x} \in \mathbb{R}^d,$$

then, the model is said to be explicit or definite. This means that the following stochastic relation is supposed to be true about the underlying structure of the data at hand:

$$Y = r_{\boldsymbol{\theta}}(\mathbf{X}) + \varepsilon,$$

where ε is the error or noise, which might or might not be distributed according to a parametric family of distributions. In this case, regression involves estimation of the parametric function $r_{\boldsymbol{\theta}}$, which is, in fact, selecting the optimal value for the parameter $\boldsymbol{\theta}$ from a predetermined hypothesis space. In contrast, if no assumption about the explicit form of the regression function $\mathbb{E}_{Y|\mathbf{X}}(y \mid \mathbf{x})$, with respect to $\mathbf{x}$, is made, then the model is called non-explicit. The well-known linear regression, for example, is an explicit model since $r_{\boldsymbol{\theta}}$ is a

polynomial in $\mathbf{X}$. On the other hand, Nadaraya–Watson kernel regression is an example of a non-explicit model [Nadaraya, 1964, Watson, 1964].

From the statistical learning point of view, in the explicit model approach one can described the solution in terms of risk minimization. Assume that there is a parametric set of admissible functions $\mathcal{H}_{\boldsymbol{\Theta}}$, where we are looking for the regression function $r_{\boldsymbol{\theta}}$. As we discussed earlier, the statistical learning theory's approach to the risk minimization problem is based on minimizing the risk over the parameter space $\boldsymbol{\Theta}$. It is easy to check that two of the most frequently used techniques of regression estimation in practice, i.e., the MLE and the *least squares*, are both special occasions of risk minimization. In fact, the difference between these two techniques lies in the losses being used. More precisely, in both cases, one desires to minimize the expected risk (3.2), once with the loss function being

$$L_{\boldsymbol{\theta}}(\mathbf{X}, Y) = -\mathcal{L}(\boldsymbol{\theta}; \mathbf{X}, Y),$$

where $-\mathcal{L}(\boldsymbol{\theta}; \mathbf{X}, Y)$ is the likelihood of $\boldsymbol{\theta}$ given $\mathbf{X}, Y$, and once with

$$L_{\boldsymbol{\theta}}(\mathbf{X}, Y) = \Big[Y - h_{\boldsymbol{\theta}}(\mathbf{X})\Big]^2,$$

i.e., the squared residual. (See, e.g., Vapnik [1998].)

Explicit modelling of regression analysis, in particular, the MLE approach is of central interest in Chapter 5. Since we, ultimately, wish to study the problem in the context of LBRC-C data, it would be insightful to become familiar with the regression tools used in survival analysis. For this reason, a brief introduction to the most frequently applied models in survival regression has been provided in subsection 3.3.6.

4.3.1 Non-Explicit Regression Estimation under LBRC-C Data

For the purpose of non-explicit modelling, it is assumed that $\mathbb{E}_{Y|\mathbf{X}}(y \mid \mathbf{x}) = r(\mathbf{x})$, where r is an unknown function. Note that we do not require r to have any particular form, such as polynomial form, for instance. What we are interested in is the value of r at any given $\mathbf{X} = \mathbf{x}_0$, i.e., $\hat{r}(\mathbf{x}_0)$, which is an estimate of $r(\mathbf{x}_0)$ calculated using the data $\{(\mathbf{x}_i, y_i) : i = 1, 2, \ldots, n\}$.

What Nadaraya-Watson estimator suggests is a *local weighted average* of values of response provided in the sample. A *kernel* is used for weighting the observations. Recall that, by definition, one needs the conditional distribution $F_{Y|\mathbf{X}}$ in order to compute the regression function. Indeed, for this purpose, the Nadaraya-Watson estimator applies the *kernel density estimation* method to estimate the conditional density $f_{Y|\mathbf{X}}$. This brings us to the following

estimator of the regression function:[2]

$$\hat{r}(\mathbf{x}_0) = \frac{\sum_{i=1}^{n} K_h(\mathbf{x}_0 - \mathbf{x}_i) y_i}{\sum_{i=1}^{n} K_h(\mathbf{x}_0 - \mathbf{x}_i)}, \tag{4.24}$$

where $K_h(\mathbf{x}) = \frac{1}{h} K(\mathbf{x})$, parameter h is the smoothing bandwidth, and K is a kernel function, i.e., a non-negative, even function with $\int K(u)\,\mathrm{d}u = 1$. This is the empirical estimator of

$$r(\mathbf{x}_0) = \frac{\mathbb{E}_{\mathbf{X},Y}\left[y\, K_h(\mathbf{x}_0 - \mathbf{x})\right]}{\mathbb{E}_{\mathbf{X}}\left[K_h(\mathbf{x}_0 - \mathbf{x})\right]}. \tag{4.25}$$

Now, let us investigate the behaviour of the estimator given in equation (4.24) in the context of LBRC-C data. The objective is to show that both the bias and censoring must be taken into account to correctly estimate the regression function. This might be achieved by applying appropriate corrections for both bias and censoring.

Given the fact that, in presence of covariates, bias itself has two levels, i.e., the length bias and the induced bias affecting the sampling distribution of the covariates, we study the effect of bias in two separate steps: First, when both biases are ignored, and second, when one fails to incorporate the covariate bias into the analysis. We do not consider the case where only length bias is ignored as it is very improbable in practice that an analyst be aware of the induced covariate bias but fails to adjust for the length bias itself. Recall that the induced covariate bias is a result of the length-biased sampling scheme, so that information on the existence of covariate bias requires the prior knowledge of length bias.

Following the same ideology as in previous sections, the discussion will be presented for Case One and Case Two, separately.

4.3.2 Regression Estimation with Pure Length Bias

In the present subsection, before introducing the proper way of estimating the regression function when data are length biased but no censoring is involved, we would like to highlight the effect of ignoring the biases resulted from the specific sampling procedure applied. As explained earlier, we will consider two distinct scenarios and evaluate the consequences of each one separately. The reason for considering these two particular scenarios is that, based on conventional approaches to regression analysis of LBRC-C data, these are, perhaps, among

[2]For simplicity, we discuss the univariate case, however, the concept remains the same for multivariate case. For instance, for multivariate covariate $\mathbf{X}$ one might apply a *product kernel*, i.e., $K_{h_1,\ldots,h_d}(x_1,\ldots,x_d) = \prod_{j=1}^{d} K_{h_j}(x_j)$.

the most probable situations that might happen in practice.

1st Naïve Approach: Ignoring Both Length Bias and Covariate Bias

Suppose that both the length bias and the induced covariate bias are ignored, and the estimator $\hat{r}(\mathbf{x}_0)$ is applied disregarding the biases. Consider, first, the numerator of equation (4.24), multiplied by the reciprocal of the sample size n: $\frac{1}{n}\sum_{i=1}^{n} K_h(\mathbf{x}_0 - \mathbf{x}_i)y_i$. Since the data are length biased, this expression, in fact, estimates $\mathbb{E}^{*}_{\mathbf{X},\tilde{Y}|\overline{T}}[y\, K_h(\mathbf{x}_0 - \mathbf{x})]$. Opening up the last expression, we have that

$$\begin{aligned}
\mathbb{E}^{*}_{\mathbf{X},\tilde{Y}|\overline{T}}\left[y\, K_h(\mathbf{x}_0 - \mathbf{x})\right] &= \int v\, K_h(\mathbf{x}_0 - \mathbf{u})\, f^{*}_{\mathbf{X},\tilde{Y}|\overline{T}}(\mathbf{u}, v)\, \mathrm{d}(\mathbf{u}, v) \\
&= \frac{1}{\mu}\int_{\mathbf{u}}\int_{v} v^2\, K_h(\mathbf{x}_0 - \mathbf{u})\, f_{Y|\mathbf{X}}(v \mid \mathbf{u})\, f_{\mathbf{X}}(\mathbf{u})\, \mathrm{d}v\, \mathrm{d}\mathbf{u} \\
&= \frac{1}{\mu}\,\mathbb{E}_{\mathbf{X},Y}\left[y^2 K_h(\mathbf{x}_0 - \mathbf{x})\right] \\
&= \frac{1}{\mu}\,\mathbb{E}_{\mathbf{X}}\left[K_h(\mathbf{x}_0 - \mathbf{x})\ \mathbb{E}_{Y|\mathbf{X}}(y^2 \mid \mathbf{X} = \mathbf{x})\right].
\end{aligned}$$

In a similar way, the denominator of equation (4.24), $\frac{1}{n}\sum_{i=1}^{n} K_h(\mathbf{x}_0 - \mathbf{x}_i)$, can be shown to equal $\mathbb{E}^{*}_{\mathbf{X}|\overline{T}}[K_h(\mathbf{x}_0 - \mathbf{x})]$, where

$$\mathbb{E}^{*}_{\mathbf{X}|\overline{T}}[K_h(\mathbf{x}_0 - \mathbf{x})] = \frac{1}{\mu}\,\mathbb{E}_{\mathbf{X}}\left[\mu(\mathbf{x})\, K_h(\mathbf{x}_0 - \mathbf{x})\right]. \tag{4.26}$$

Now, if $\hat{r}_{\text{naïve}}(\mathbf{x}_0)$ denotes the naïve estimator inspired by equation (4.24), i.e.,

$$\hat{r}_{\text{naïve}}(\mathbf{x}_0) := \frac{\sum_{i=1}^{n} K_h(\mathbf{x}_0 - \mathbf{x}_i)y_i}{\sum_{i=1}^{n} K_h(\mathbf{x}_0 - \mathbf{x}_i)}, \tag{4.27}$$

then, in fact, instead of $r(\mathbf{x}_0)$ one estimates

$$\frac{\mathbb{E}_{\mathbf{X},Y}\left[y^2 K_h(\mathbf{x}_0 - \mathbf{x})\right]}{\mathbb{E}_{\mathbf{X}}\left[\mu(\mathbf{x}) K_h(\mathbf{x}_0 - \mathbf{x})\right]},$$

which is, asymptotically, equal to

$$\frac{\mathbb{E}_{Y|\mathbf{X}}\left[y^2 \mid \mathbf{X} = \mathbf{x}_0\right]}{\mathbb{E}_{Y|\mathbf{X}}\left[y \mid \mathbf{X} = \mathbf{x}_0\right]}.$$

2nd Naïve Approach: Ignoring Covariate Bias

The next scenario to be considered is when takes into account the length bias correctly, but ignores the induced covariate bias. Note that this is, perhaps, the most common mistake that can occur in practice.

Taking care of the length bias but ignoring the covariate bias at the same time means that in the numerator of equation (4.25), i.e.,

$$\mathbb{E}_{\mathbf{X},Y}\left[y\, K_{h}(\mathbf{x}_0 - \mathbf{x})\right] = \int_{\mathbf{u}} \int_{v} v\, K_{h}(\mathbf{x}_0 - \mathbf{u})\, f_{Y|\mathbf{X}}(v \mid \mathbf{u}) f_{\mathbf{X}}(\mathbf{u})\, \mathrm{d}v\, \mathrm{d}\mathbf{u}$$

$f_{Y|\mathbf{X}}$ has been correctly plugged in by properly reweighing the observed length-biased conditional distribution $f_{\tilde{Y}|\mathbf{X}}^{*}$, however, instead of $f_{\mathbf{X}}$, mistakenly, $f_{\mathbf{X}|T}^{*}$ has been used. This yields the following:

$$\begin{aligned}\int_{\mathbf{u}} \int_{v} v K_{h}(\mathbf{x}_0 - \mathbf{u})\, f_{Y|\mathbf{X}}(v \mid \mathbf{u})\, \frac{\mu(\mathbf{u})}{\mu} f_{\mathbf{X}}(\mathbf{u})\, \mathrm{d}v\, \mathrm{d}\mathbf{u} \\ = \frac{1}{\mu} \int_{\mathbf{u}} \left[\int_{v} v\, f_{Y|\mathbf{X}}(v \mid \mathbf{u})\, \mathrm{d}v\right] \mu(\mathbf{u}) K_{h}(\mathbf{x}_0 - \mathbf{u})\, f_{\mathbf{X}}(\mathbf{u})\, \mathrm{d}\mathbf{u} \\ = \frac{1}{\mu}\, \mathbb{E}_{\mathbf{X}}\left[\mu(\mathbf{x})^2 K_{h}(\mathbf{x}_0 - \mathbf{x})\right].\end{aligned}$$

The denominator remains as in the previous case, expressed in equation (4.26). Consequently, applying the naïve estimator $\hat{r}_{\text{naïve}}(\mathbf{x}_0)$, defined by equation (4.27), results in the estimation of

$$\frac{\mathbb{E}_{\mathbf{X}}\left[\mu(\mathbf{x})^2\, K_{h}(\mathbf{x}_0 - \mathbf{x})\right]}{\mathbb{E}_{\mathbf{X}}\left[\mu(\mathbf{x})\, K_{h}(\mathbf{x}_0 - \mathbf{x})\right]},$$

instead of $r(\mathbf{x}_0)$.

Proper Approach: The Corrected Estimator

To correct the estimation, one needs to consider the following estimator that can be easily shown to be the right choice:

$$\hat{r}_{\text{LB}}(\mathbf{x}_0) = \frac{\sum_{i=1}^{n} K_{h}(\mathbf{x}_0 - \mathbf{x}_i)}{\sum_{i=1}^{n} y_i^{-1} K_{h}(\mathbf{x}_0 - \mathbf{x}_i)}.$$

4.3.3 Regression Estimation with Length Bias and Right Censoring

When data are right censored, the situation is different since, apparently, for part of the data the information about lifetime is only partially available. Nonetheless, we will show that, surprisingly interesting, one can rely *only* on the failed portion of the data in order to estimate the regression function even when some individuals are right censored. This is because, first, based on equation (4.25), for computing $r(\mathbf{x}_0)$ the unbiased joint distribution of $\mathbf{X}, Y$ is needed. On the other hand, equation (4.14) illustrated that $F_{\mathbf{X},Y}$ may be expressed, solely, by the biased joint distribution of the failed subjects $\tilde{F}_1$.

However, one needs the survival function of the residual censoring C since as shown in equation (4.14), $\tilde{F}_1$ also depends on the residual censoring distribution. Recall that index 1 represented the failure, i.e., $\delta = 1$. A slight modification of equation (4.13), provides that

$$f^{*}_{\mathbf{X},\tilde{Y},\delta|\overline{T}}(\mathbf{x}, y, \delta = 1) = \frac{f_{\mathbf{X},Y}(\mathbf{x}, y)}{\mu}\,\zeta(y),$$

where $\zeta(y)$ was defined to be a function of the survival function of the residual censoring time, i.e.,

$$\zeta(y) = \int_0^y S_C(r)\,\mathrm{d}r = \int_0^y \mathbb{P}(C > r)\,\mathrm{d}r.$$

Therefore, one can write that

$$\begin{aligned}
\mathbb{E}_{\mathbf{X},Y}\left[y\,K_h(\mathbf{x}_0 - \mathbf{x})\right] &= \int_{\mathbf{u},v} v\,K_h(\mathbf{x}_0 - \mathbf{u})\,\mathrm{d}F_{\mathbf{X},Y}(\mathbf{u}, v) \\
&= \mu \int_{\mathbf{u},v} \frac{v}{\zeta(v)}\,K_h(\mathbf{x}_0 - \mathbf{u})\,f^{*}_{\mathbf{X},\tilde{Y},\delta|\overline{T}}(\mathbf{u}, v, \delta = 1)\,\mathrm{d}(\mathbf{u}, v) \\
&= \mu\,\mathbb{E}^{*}_{\mathbf{X},\tilde{Y},\delta|\overline{T}}\left[\frac{y}{\zeta(y)}K_h(\mathbf{x}_0 - \mathbf{x})\,\mathbb{1}(\delta = 1)\right].
\end{aligned}$$

In a similar way, one can easily show that

$$\mathbb{E}_{\mathbf{X}}[K_h(\mathbf{x}_0 - \mathbf{x})] = \mathbb{E}_{\mathbf{X},Y}[K_h(\mathbf{x}_0 - \mathbf{x})] = \mu\,\mathbb{E}^{*}_{\mathbf{X},\tilde{Y},\delta|\overline{T}}\left[\frac{1}{\zeta(y)}K_h(\mathbf{x}_0 - \mathbf{x})\,\mathbb{1}(\delta = 1)\right].$$

The last two equations suggest that the following estimator can actually estimate $r(\mathbf{x}_0)$ correctly:

$$\hat{r}_{\text{LBRC}}(\mathbf{x}_0) = \frac{\sum_{i:\delta_i=1} y_i[\zeta(y_i)]^{-1}K_h(\mathbf{x}_0 - \mathbf{x}_i)}{\sum_{i:\delta_i=1}[\zeta(y_i)]^{-1}K_h(\mathbf{x}_0 - \mathbf{x}_i)}.$$

CHAPTER 5 VARIABLE SELECTION IN LBRC-C REGRESSION SETTING

The problem of *variable* or *feature selection* in both statistics and machine learning is of significant importance from the inferential as well as the computational perspectives [Sauer et al., 2013, Chowdhury and Turin, 2020, Genuer et al., 2010a, Meyer et al., 2019]. Moreover, from the inferential point of view, both of the so-called descriptive and predictive models might benefit from variable selection, considerably. Also, variable selection might be applied in both classification and regression problems. Examples of incentives to select a subset of variables include (i) to detect the main risk factors responsible for developing a health condition, (ii) to initiate appropriate preventive measures to avoid adverse outcomes of a treatment, (iii) to reduce the variation of predictions due to rather small sample size compared to the number of features, (iv) to gain interpretability, (v) to gain computational efficiency and to reduce the training and testing time, (vi) to avoid obsolete model complexity, and consequently, to reduce the chance of overfitting. Especially, due to new data collecting technologies, high-dimensional data become more prevalent in numerous areas, in which case variable selection may play a crucial role [Zhang et al., 2008]. For different applications of variable selection, one may check Akarachantachote et al. [2014], Genuer et al. [2010b], Lu and Petkova [2014], Meyer et al. [2019], among others. The aforementioned examples are some of the most important purposes due to which variable selection might be performed. Nonetheless, reasons for variable selection are not restricted to the items on this list.

5.1 Conditional and Unconditional Approaches to Variable Selection

The core topic of the current chapter is variable selection in the context of explicit regression analysis (defined in Section 4.3) of LBRC-C data. More precisely, two methods of variable selection in the context of length-biased, and right-censored time-to-event data are considered and their properties in terms of selecting the *correct* model, defined later, are investigated in detail. Both of the methods are based on the MLE.

A key aspect of the discussion in this chapter is that it was motivated by the classical approaches to the regression analysis of time-to-event data, especially, the so-called AFT models. Hence, modelling the survival function, or equivalently the distribution function, is the center of attention. Additionally, due to the parametric nature of the analysis, it is assumed that the response distribution belongs to a priori known family of parametric distributions. Nonetheless, using the techniques provided in the previous chapter many of

the results obtained in the present chapter can be modified to suit the non-parametric or semi-parametric frameworks.

Once the parametric paradigm is considered, MLE becomes our central inferential machinery Vapnik [1998]. The MLE is truly considered as the essence of inference in parametric statistics for its distinctively excellent conceptual and technical properties, which can be checked in many classical texts [Wilks, 1943, Lehmann, 1999, 1983, Lehmann and Casella, 1998, Casella and Berger, 2002]. Besides its inferential power, it provides a very intuitive and naturally justifiable tool for analysis in the parametric framework. To see this, one simply needs to analyze and to understand the definition of the likelihood function. Nevertheless, it will be shown that in case of length-biased sampling, basing estimation on likelihood requires paying extra attention to the covariate bias induced by the sampling scheme.

Now, let us elaborate on the two approaches we consider. One, which is referred to as the *conventional* approach, involves *conditioning* the likelihood function on the distribution of the covariates, while the second employs the joint likelihood of the covariates and the response. This second approach will be called the *joint* approach. These two criteria, differ in one important respect as follows: The joint likelihood function fuses, properly, with the covariate bias, whereas the conditional one fails to take this effect into account. While the joint approach was initially proposed, by Bergeron [2006] and Bergeron et al. [2008], to estimate the parameters in regression analysis of LBRC-C data, here in this work, we generalize it to be applied to the problem of variable selection when data are LBRC-C.

It turns out that the question of the potential effects of the covariate bias had not received much attention prior to Bergeron et al. [2008]. In fact, Begg and Gray [1987] seem to be the only authors accounting for the covariate bias in the estimation of odds ratio in a prevalent cohort case-control study. Nevertheless, the problem of parameter estimation as well as variable selection, classification, and the rest of the problems considered in our research were not studied, previously.

In terms of parameter estimation, Bergeron [2006] and Bergeron et al. [2008] thoroughly investigated the implications of estimating parameters by means of the joint approach and derived its distinctions with the conventional one. In particular, they showed that with left truncation, estimating the parameter based on the conditional likelihood yields biased estimations since it ignores the information carried by the biased covariates. Moreover, they established that, in contrast, grounding the analysis in its joint counterpart incorporates this information into the estimation and produces superior estimations. It appeared that the conditional likelihood, besides generating biased estimations, negatively affects the *efficiency* of the estimation procedure. The *loss of efficiency* becomes even more problematic when

the sample size is small. In practice, there are countless occasions where training data is extremely limited and, therefore, managing the resources at disposal turns to be crucial.

For the first time in the literature, the present work provides a novel method of selection that generalizes the application of the joint likelihood to the feature selection problem in the setting of interest, i.e., LBRC-C data. It is worth mentioning that besides variable selection and regression, in general, classification of lifetimes in presence of length bias and right censoring has not been studied in the literature either. Here, while an in-depth study variable selection is provided, the classification problem is left for future research. In the context of variable selection, our research, specifically, concludes that the joint approach is the superior method, compared to the conventional one, in selecting the *correct model*, in the sense explained later. The technical details are provided in the remainder of this chapter.

The final section of this chapter draws a comparison between the two approaches by a brief simulation study, which confirms the theoretical expectations, especially, with small samples. This section highlights the use of information criteria constructed based on the likelihoods discussed earlier for selecting the optimal subset of the original variables. Particularly, special attention is paid to the well-known *AIC* and *BIC*, introduced by Akaike [1974] and [Schwarz, 1978], respectively. The reason is that, firstly, both of these criteria are considered mainstream choices for *model selection* (including variable selection), and hence, are very popular. Secondly, they constitute a major base for many other criteria which were developed later in order to improve the performance of the AIC and BIC in specific respects.

Despite being popular in practice, applying such criteria, like the AIC and BIC, swiftly becomes expensive and inefficient as the dimension of the covariate vector increases. The reason is that, as these criteria were initially created for model selection, they do not select variable at the same time as they estimate the parameters, which means they must be iteratively applied to each subset of the original variables, separately, in order to estimate the parameters first, and only then different models, which are different subsets of features, might be compared according to their likelihoods. Therefore, for higher dimensions, the direction of focus should be switched towards methods that perform the estimation and variable selection simultaneously. Examples of such criteria include the *LASSO* [Santosa and Symes, 1986, Tibshirani, 1996], the *Adaptive LASSO* [Zou, 2006], the *Smoothly Clipped Absolute Deviation (SCAD)* [Fan and Li, 2001], the *Group LASSO* [Yuan and Lin, 2006], and the *Elastic Net* method introduced by [Zou and Hastie, 2005].

Before, formalizing the problem of variable selection in the next section, let us briefly introduce the conditional and the joint likelihood. The conditional approach is, in fact, the one which is predominantly used in the regression analysis of survival time. The core part of this

approach is the fact that the parameter estimation is performed utilizing the conditional distribution of the response Y_i, conditioned on the covariate $\mathbf{X}_i$. Let $\mathcal{L}_I$ denote the conditional likelihood function[1] and $\{(\mathbf{x}_i, y_i) : i = 1, 2, \ldots, n\}$ be a realized sample dataset. Then,

$$\mathcal{L}_I(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i \mid \mathbf{x}_i; \boldsymbol{\theta}),$$

where f is a generic density and $\boldsymbol{\theta}$, the vector of free parameters. Similarly, one defines the joint likelihood based on the joint distribution of Y_i and $\mathbf{X}_i$ as follows:

$$\mathcal{L}_J(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(\mathbf{x}_i, y_i; \boldsymbol{\theta}).$$

Note that these definitions are general and does not depend on specific distribution of data. In the following section, the problem of variable selection is formalized in detail.

5.2 General Statement of the Variable Selection Problem

Consider the regression equation

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_d X_d + \varepsilon, \tag{5.1}$$

where $Y \in \mathbb{R}$ is the response, $\mathbf{X} = (X_1, \ldots, X_d) \in \mathbb{R}^d$ is a vector of covariates, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_d) \in \mathcal{B} \subset \mathbb{R}^{d+1}$ denotes the regression coefficients, and ε represents a suitable error term independent from $\mathbf{X}$. Hence, the regression function $r_{\boldsymbol{\beta}}$, which is a real-valued linear function of $\boldsymbol{\beta}$ and $\mathbf{X}$, has an explicit parametric form.

$$r_{\boldsymbol{\beta}}(\mathbf{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_d X_d.$$

Suppose that the sample data provide a complete set of d covariates for each subject, i.e., a vector $\mathbf{X}_i = (X_{i1}, \ldots, X_{id})$ is associated with each subject i. Here, the sample size is usually assumed to be n, unless otherwise is specified. The regression model specified by equation (5.1) is called a *full model* if none of $X_1, \ldots, X_d$ is excluded form the model. In a certain sense, the full model is the most complex model one could possibly build with the data in hand.[2]

[1] The subscript I in $\mathcal{L}_I$ is for *ignorance*.

[2] Here, by complexity we mean number of the variables a model contains. This should not be confused with the concept of complexity in VC theory.

Assuming one-to-one correspondence between the set of functions $r_{\boldsymbol{\beta}}$ and the parameter space $\mathcal{B}$, function $r_{\boldsymbol{\beta}}$ can be represented uniquely by its parameter, in which case the problem of variable selection could be formalized as follows: Denote $\mathcal{I} = \{1, 2, \ldots, d\}$ and let $\boldsymbol{\beta}^0 = ({\boldsymbol{\beta}^0}_0, \ldots, {\boldsymbol{\beta}^0}_d)$ be the true regression coefficient vector, some of whose entries may possibly be zero. Recall that $\mathcal{B} \subset \mathbb{R}^{d+1}$ is the coefficient space. For any $\boldsymbol{\beta} \in \mathcal{B}$ define

$$\mathcal{N}(\boldsymbol{\beta}) := \{j \in \mathcal{I} | \, \beta_j \neq 0\}.$$

In other words, $\mathcal{N}$ extracts indices of non-zero entries of each coefficient vector $\boldsymbol{\beta}$, regardless of its intercept β_0. For example, if $\boldsymbol{\beta} = (3, 0, -8.1, 7.33)$, then, $d = 3$, $\mathcal{I} = \{1, 2, 3\}$, and $\mathcal{N}(\boldsymbol{\beta}) = \{2, 3\}$.

It is easy to see that $\mathcal{N}(\boldsymbol{\beta})$ provides a partition of $\mathcal{I}$, i.e., $\mathcal{N}(\boldsymbol{\beta})$ and its complement $\mathcal{I} \backslash \mathcal{N}(\boldsymbol{\beta})$, for any $\boldsymbol{\beta} \in \mathcal{B}$. Equivalently, $\mathcal{N}$ partitions $\mathcal{B}$ into 2^d equivalence classes according to zero and non-zero indices of coefficients. That is, for any $\boldsymbol{\alpha} \in \mathcal{B}$, its equivalence class is

$$\mathcal{B}_{\boldsymbol{\alpha}} := [\boldsymbol{\alpha}] = \{\boldsymbol{\beta} \in \mathcal{B} : \mathcal{N}(\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\alpha})\}. \tag{5.2}$$

Excluding the trivial, and obviously, uninteresting class of $\mathcal{B}_{\mathbf{0}}$, where $\mathbf{0}$ denotes the d-dimensional zero vector, leaves one with $2^d - 1$ classes of coefficients.

Variable (feature) selection in the context of regression problem (5.1) involves *finding* $\mathcal{N}(\boldsymbol{\beta}^0)$ *from the provided set of sample data*. There is an important point worth mentioning here: In the variable-selection literature, different criteria have been used in order to measure how well a selection procedure performs. What criterion to pick for evaluation must be decided based on the context and objective of the regression analysis. Further, we introduce a couple of these criteria that are commonly used in variable selection problems.

The first set of criteria is motivated by prediction ability of a model when new input vectors are given. Therefore, in this case the *generalizability* of a model is emphasized. Accordingly, a model $\boldsymbol{\beta} \in \mathcal{B}$ is said to be an *overfitted* model if $\mathcal{N}(\boldsymbol{\beta}^0) \subset \mathcal{N}(\boldsymbol{\beta})$. That is, if an entry of the overfitted $\boldsymbol{\beta}^0$ differs from zero, then the corresponding entry of $\boldsymbol{\beta}$ is non-zero too. Similarly, $\boldsymbol{\beta}$ is called a *fitted* model if $\mathcal{N}(\boldsymbol{\beta}^0) = \mathcal{N}(\boldsymbol{\beta})$, i.e., if $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^0$ belong to the same equivalence class of $\mathcal{B}$ induced by $\mathcal{N}$. And lastly, $\boldsymbol{\beta}$ will be called *underfitted* if it is neither overfitted nor fitted, i.e., there is at least one non-zero entry of $\boldsymbol{\beta}^0$ which has been set to zero in $\boldsymbol{\beta}$ [Guyon and Yao, 1999, Aghababaei Jazi, 2019, Bozdogan and Haughton, 1998].

Later in this chapter, it will be demonstrated that variable selection based on $\mathcal{L}_J$ is, consistently, less susceptible to underfitting compared to the selection based on $\mathcal{L}_I$. This aspect of the difference between the two procedures becomes more meaningful when the objective

of the analysis is, e.g., to detect the true risk factors related to a condition or disease. This is the setting where prediction of the survival time is not the main objective of the analysis, but rather recognition of possibly influential factors on lifetime.

Shao [1993] provides another division of the parameter space, where $\mathcal{B}$ is divided into two partitions of *correct* and *incorrect* models: Any $\boldsymbol{\beta} \in \mathcal{B}$ that satisfies the inclusion $\mathcal{N}(\boldsymbol{\beta}^0) \subseteq \mathcal{N}(\boldsymbol{\beta})$ is called a correct model. This is clearly equivalent to the union of the fitted and overfitted models above. Subsequently, any model that is not a correct model is regarded as *incorrect*, i.e., the set of underfitted models. In this setting, whose main concern is apparently *not to miss out on influential factors*, our results on the difference between the conditional and the joint approaches become even more highlighted since $\mathcal{L}_I$ has a stronger tendency to select incorrect models compared to $\mathcal{L}_J$.

5.3 Likelihood-Based Selection Procedure

As mentioned earlier, variable selection might be considered as a special case of model selection. This means that one may use model selection criteria for variable selection as well. For example, the information-based, model selection criteria, such as the AIC, the BIC, as well as their variants can, also, be applied to select a subset of the available variables. On the other hand, there are criteria that have been developed specifically for variable selection, such as the LASSO. In addition, both the variable and model selection criteria may be constructed based on the MLE [Fan and Peng, 2004]. However, it is important to notice that the selection procedure is different depending on whether a model or variable selection criterion is used. In what follows, this difference is indicated, briefly.

A likelihood-based model selection criterion is, usually, comprised of two components: (i) The so-called *goodness-of-fit* term, which is based on the (log-)likelihood function or a function of it, and (ii) a regularization term, which is essentially a *penalty* on the model *complexity*. Therefore, a generic model selection principle $\mathcal{M}$, based on likelihood estimation is of the following general form:

$$\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}: \quad \mathcal{C}\left(\boldsymbol{\theta}; \mathcal{D}\right) := L\left(\hat{\boldsymbol{\theta}}_{\text{MLE}}\right) - \boldsymbol{P}(\eta, |\boldsymbol{\theta}|), \tag{5.3}$$

where $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ is the MLE of $\boldsymbol{\theta}$, L is a function of the likelihood function, $\boldsymbol{P}(\eta, |\boldsymbol{\theta}|)$ a penalty term, which may or may not depend on the sample size n, and $|\boldsymbol{\theta}|$ represents the model complexity, which is usually the dimension of the parameter $\boldsymbol{\theta}$. Note that, instead of $\boldsymbol{\beta}$, the model is represented by $\boldsymbol{\theta}$ to emphasize that it does not necessarily need to be the regression coefficient.

When a set of candidate models $\Delta \subset \boldsymbol{\Theta}$ is given, model selection employing the measure (5.3) is equivalent to solving the following optimization problem:

$$\arg\max_{\boldsymbol{\theta}\in\Delta}\{\mathcal{C}(\boldsymbol{\theta};\mathcal{D})\}. \tag{5.4}$$

When such a criterion is used for variable selection, then each subset of the original d variables constitute a model that corresponds to a certain equivalent class $\mathcal{B}_{\boldsymbol{\beta}}$, defined by (5.2). After solving problem (5.4), the equivalence class corresponding to the solution of (5.4) is used to extract the non-zero coefficients. As one can see, this is a tedious procedure as solving the optimization problem involves calculating $\mathcal{C}$ for every member of Δ. This is the reason for our earlier claim that as the number of covariates increases, using information-based model selection criteria, such as AIC or BIC, for variable selection becomes inefficient very quickly.

On the other hand, variable-selection-specific criteria perform the estimation and variable selection at the same time, as a result of which, the necessity of iterating the calculation for each individual subset of variables is eliminated. Variable selection criteria, also, posses two components, i.e.,

$$\mathcal{C}(\boldsymbol{\beta};\mathcal{D}) := \ell(\boldsymbol{\beta};\mathcal{D}) - \sum_{j=1}^{d}\mathcal{P}_{\eta}(|\beta_j|),$$

where $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_d)$, $\ell(\boldsymbol{\beta};\mathcal{D})$ is the log-likelihood of parameter $\boldsymbol{\beta}$ given data $\mathcal{D} = \{\mathbf{z}_i : i = 1,2,\ldots,n\}$, $\mathcal{P}_{\eta}$ is a penalty function, possibly depending on n, and $\eta > 0$ is a tuning parameter [Fan and Peng, 2004]. Here, we also solve an optimization problem defined as

$$\arg\max_{\boldsymbol{\beta}\in\mathcal{B}}\{\mathcal{C}(\boldsymbol{\beta};\mathcal{D})\}.$$

The criterion is fed by the full set of variables, immediately, but the penalty term is designed in a way that some of the coefficients are forcibly shrank to zero, during the estimation step, given the tuning parameter is tuned properly. Variable selection criteria has been studied and fully explained in numerous articles and textbooks, so we skip further details in this work. Interested reader may refer to James et al. [2013], Hastie et al. [2009], Desboulets [2018].

5.4 The Two Likelihoods and their Discrepancy Due to Bias

This section is devoted to a more detailed analysis of the conditional and the joint likelihoods $\mathcal{L}_I$ and $\mathcal{L}_J$. Particularly, we discuss the differences between the two and explain the problem resulted from applying the wrong likelihood, i.e., the conditional one, in regression analysis of LBRC-C data. Eventually, these distinctions will be used to show that this is an issue

affecting *not* only parameter estimation, as correctly pointed out by Bergeron et al. [2008], Bergeron [2006], but also spreads to variable selection (and presumably, classification) if analysis is base on the MLE. First, the exact forms of $\mathcal{L}_I$ and $\mathcal{L}_J$ are derived. These are the specific forms resulted from the data being length biased and right censored.

5.4.1 Derivation of the Conditional and Unconditional Likelihoods

Let us start from the case where there are no covariates in the sample data, i.e., the available data are of the following form: $\{(\tilde{A}_i, \tilde{R}_i \wedge C_i, \delta_i) : i = 1, \ldots, n\}$. Vardi [1989] derived the likelihood under multiplicative censoring as follows:

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{ \left[\frac{y_i f_Y(y_i; \boldsymbol{\theta})}{\mu(\boldsymbol{\theta})} \right]^{\delta_i} \left[\int_{v \geq a_i + c_i} \frac{v f_Y(v; \boldsymbol{\theta})}{\mu(\boldsymbol{\theta})} \, \mathrm{d}v \right]^{1-\delta_i} \right\}, \tag{5.5}$$

where $\mu(\boldsymbol{\theta}) = \mathbb{E}_Y(y)$, i.e., the overall mean lifetime in the incident population. Vardi described multiplicative sampling as follows. Let $\tilde{Y}_1, \ldots, \tilde{Y}_{n_1+n_2}$ be identically distributed random variables and $\eta_1, \ldots, \eta_{n_1} \overset{\text{i.i.d.}}{\sim} \mathrm{U}(0,1)$. Then, we observe $\xi_1, \ldots, \xi_{n_1}, \tilde{Y}_{n_1+1}, \ldots, \tilde{Y}_{n_1+n_2}$, with $\xi_i = \eta_i \tilde{Y}_i$, $i = 1, \ldots, n_1$. In cross-sectional sampling context, this is equivalent to recruiting $n_1 + n_2$ individuals into the study and censoring n_1 of them immediately. On the other hand, a more flexible approach would let all individuals remain in the study until they have either failed or been censored. In fact, in contrast to the multiplicative censoring, this allows for a random number of censored and failed subjects, i.e., n_c and n_f, respectively. This is a more general setting and as mentioned in subsection 4.1.1, is the setting we have adopted here.

It is known that length-biased sampling causes informative censoring; likewise, is multiplicative censoring as the censoring and survival times are clearly dependent, as shown in subsection 4.1.1. Although, Vardi's likelihood (5.5) correctly accounts for the informativeness of censoring, it cannot be incautiously extended to the case with covariates, as we will see shortly.

Now, let a vector of covariates $\overset{*}{\mathbf{X}}_i$ is also provided for each data point in the sample, i.e., $\{(\overset{*}{\mathbf{X}}_i, \tilde{A}_i, \tilde{R}_i \wedge C_i, \delta_i) : i = 1, \ldots, n\}$. As mentioned earlier, in regression analysis of LBRC-C data, one would conventionally condition the likelihood function on $\overset{*}{\mathbf{X}}_i$. A naïve extension

of Vardi's likelihood to the new case, where covariates are also present, is as follows then:

$$\mathcal{L}_{I}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{ \left[\frac{y_i f_{Y}(y_i \mid \mathbf{x}_i; \boldsymbol{\theta})}{\mu(\mathbf{x}_i; \boldsymbol{\theta})} \right]^{\delta_i} \left[\int\limits_{v \geq a_i + c_i} \frac{v f_{Y}(v \mid \mathbf{x}_i; \boldsymbol{\theta})}{\mu(\mathbf{x}_i; \boldsymbol{\theta})} \, \mathrm{d}v \right]^{1-\delta_i} \right\}, \tag{5.6}$$

where $\mu(\mathbf{x}_i; \boldsymbol{\theta}) = \mathbb{E}_{Y|\mathbf{X}}(y \mid \mathbf{X} = \mathbf{x}_i)$. The likelihood (5.6) seems to be a natural extension of Vardi's likelihood (5.5). However, the problem is that it ignores the information provided by the sampling distribution of the covariates. This is, as claimed multiple times before, is the main source of the problem. To understand this, we need to recall a fact, which was discussed at the beginning of Chapter 3: While the generator assigns responses according to the conditional distribution of the response, given the covariate, what is observed by the learning machine is a set of examples coming from the joint distribution. This is aligned with the form of each observation stated above, i.e., $(\overset{*}{\mathbf{X}}_i, \tilde{A}_i, \tilde{R}_i \wedge C_i, \delta_i)$.

Therefore, it must be easy to see why one should use the joint likelihood, and not the conditional one, in the first place. Now, one may wonder how this simple fact could possibly be ignored in the regression analysis of LBRC-C data. As matter of fact, when there is no covariate bias, as long as maximizing the likelihood function is concerned, the conditional and joint likelihoods are equivalent, optimization-wise. That is, maximizing the conditional likelihood provides the same result as maximizing the joint one. There is a subtle difference between the unbiased and the biased cases that leads to this confusion. As we will clarify in the following passages, when data are collected through the prevalent-cohort, cross-sectional sampling design, due to the induced covariate bias, the two likelihoods are no longer equivalent.

As the first step, check that the joint likelihood, based on observations, is

$$\begin{aligned}
\mathcal{L}_{J}(\boldsymbol{\theta}) :&= \prod_{i=1}^{n} f^{*}_{\mathbf{X},\tilde{A},\tilde{R}\wedge C,\delta|\overline{T}}(\mathbf{x}_i, a_i, r_i \wedge c_i, \delta_i; \boldsymbol{\theta}) \\
&= \prod_{i=1}^{n} \left[f^{*}_{\mathbf{X},\tilde{Y}|\overline{T}}(\mathbf{x}_i, y_i; \boldsymbol{\theta}) \right]^{\delta_i} \left[\int\limits_{v \geq a_i + c_i} f^{*}_{\mathbf{X},\tilde{Y}|\overline{T}}(\mathbf{x}_i, v; \boldsymbol{\theta}) \, \mathrm{d}v \right]^{1-\delta_i} \\
&= \prod_{i=1}^{n} \left[f_{\tilde{Y}|\overset{*}{\mathbf{X}},\overline{T}}(y_i \mid \mathbf{x}_i; \boldsymbol{\theta}) \right]^{\delta_i} \left[\int\limits_{v \geq a_i + c_i} f_{\tilde{Y}|\overset{*}{\mathbf{X}},\overline{T}}(v \mid \mathbf{x}_i; \boldsymbol{\theta}) \, \mathrm{d}v \right]^{1-\delta_i} f^{*}_{\mathbf{X}|\overline{T}}(\mathbf{x}_i; \boldsymbol{\theta}) \\
&= \prod_{i=1}^{n} \left[\frac{y_i f_{Y|\mathbf{X}}(y_i \mid \mathbf{x}_i; \boldsymbol{\theta})}{\mu(\mathbf{x}_i; \boldsymbol{\theta})} \right]^{\delta_i} \left[\int\limits_{v \geq a_i + c_i} \frac{v f_{Y|\mathbf{X}}(v \mid \mathbf{x}_i; \boldsymbol{\theta})}{\mu(\mathbf{x}_i; \boldsymbol{\theta})} \, \mathrm{d}v \right]^{1-\delta_i} f^{*}_{\mathbf{X}|\overline{T}}(\mathbf{x}_i; \boldsymbol{\theta}).
\end{aligned} \tag{5.7}$$

Now, equations (5.6) and (5.7) imply

$$\mathcal{L}_{J}(\boldsymbol{\theta}) = \mathcal{L}_{I}(\boldsymbol{\theta}) \prod_{i=1}^{n} f^{*}_{\mathbf{X}|\overline{T}}(\mathbf{x}_{i};\boldsymbol{\theta}), \tag{5.8}$$

where $f^{*}_{\mathbf{X}|\overline{T}}(\mathbf{x}_{i};\boldsymbol{\theta})$ is the density of the biased covariate. Clearly, $\mathcal{L}_{J}$ is proportional to $\mathcal{L}_{I}$, however, due to the induced covariate bias, $f^{*}_{\mathbf{X}|\overline{T}}(\mathbf{x}_{i};\boldsymbol{\theta})$ is informative, i.e., it depends on the parameter $\boldsymbol{\theta}$. When there is no bias, this last term does not depend on the parameter $\boldsymbol{\theta}$ and maximizing either of $\mathcal{L}_{I}$ or $\mathcal{L}_{J}$ provides the same result. In fact, for unbiased data, say $\mathcal{D} = \{(\mathbf{X}_{i}, A_{i}, R_{i} \wedge C_{i}, \delta_{i} : i = 1, 2, \ldots, n\}$, holds the following:

$$\mathcal{L}_{J}(\boldsymbol{\theta};\mathcal{D}) = \mathcal{L}_{I}(\boldsymbol{\theta};\mathcal{D}) \prod_{i=1}^{n} f_{\mathbf{X}}(\mathbf{x}_{i}),$$

where the $\mathbf{X}$ does not depend on $\boldsymbol{\theta}$. This might be not readily obvious why the distribution of the covariates $\overset{*}{\mathbf{X}}$ is dependent on the parameter. Intuitively, this happens because longer survivors, which are favored by the length-biased sampling procedure, bring in with themselves those values of the covariates, which are highly associated with longer lifetimes.

Apart from this intuitive explanation, one can mathematically demonstrate the dependence of $f^{*}_{\mathbf{X}}$ on $\boldsymbol{\theta}$ as follows: Note that be definition, we have that

$$\begin{aligned} f^{*}_{\mathbf{X}|\overline{T}}(\mathbf{x};\boldsymbol{\theta}) &= f(\mathbf{X} = \mathbf{x} \mid \overline{T};\boldsymbol{\theta}) \\ &= \frac{\mathbb{P}(\overline{T} \mid \mathbf{X} = \mathbf{x};\boldsymbol{\theta}) f(\mathbf{X} = \mathbf{x};\boldsymbol{\theta})}{\mathbb{P}(\overline{T};\boldsymbol{\theta})}, \end{aligned}$$

where f denotes a generic distribution. Note that the second term in the numerator, i.e., $f(\mathbf{X} = \mathbf{x};\boldsymbol{\theta})$, is free from truncation $\overline{T}$. In other words, it refers to the covariate distribution in the incident population, and consequently, does not depend on $\boldsymbol{\theta}$ either. Hence,

$$f^{*}_{\mathbf{X}|\overline{T}}(\mathbf{x};\boldsymbol{\theta}) = \frac{\mathbb{P}(\overline{T} \mid \mathbf{X} = \mathbf{x};\boldsymbol{\theta})\, f_{\mathbf{X}}(\mathbf{x})}{\mathbb{P}(\overline{T};\boldsymbol{\theta})}.$$

Now, applying the law of total probability to the denominator, one obtains that

$$f^{*}_{\mathbf{X}|\overline{T}}(\mathbf{x};\boldsymbol{\theta}) = \frac{\mathbb{P}(\overline{T} \mid \mathbf{X} = \mathbf{x};\boldsymbol{\theta}) f_{\mathbf{X}}(\mathbf{x})}{\displaystyle\int \mathbb{P}(\overline{T} \mid \mathbf{u};\boldsymbol{\theta}) f_{\mathbf{X}}(\mathbf{u})\, \mathrm{d}\mathbf{u}}.$$

Recall that $\overline{T}$ represents the event $Y \geq A$. Let τ be a large constant that covers the range of possible lifetimes in the target population. Since Y and A are independent and A is uniformly

distributed (stationarity assumption), one can check that

$$\mathbb{P}(\overline{T} \mid \mathbf{X} = \mathbf{x}; \boldsymbol{\theta}) = \frac{1}{\tau}\,\mu(\mathbf{x}; \boldsymbol{\theta}).$$

Now, plugging the last expression in equation (5.9) yields

$$\begin{aligned} f^{*}_{\mathbf{X}|\overline{T}}(\mathbf{x}; \boldsymbol{\theta}) &= \frac{\frac{1}{\tau}\mu(\mathbf{x}; \boldsymbol{\theta}) f_{\mathbf{X}}(\mathbf{x})}{\frac{1}{\tau}\int \mu(\mathbf{u}; \boldsymbol{\theta}) f_{\mathbf{X}}(\mathbf{u})\,\mathrm{d}\mathbf{u}} \\ &= \frac{\mu(\mathbf{x}; \boldsymbol{\theta}) f_{\mathbf{X}}(\mathbf{x})}{\mu(\boldsymbol{\theta})}, \end{aligned} \tag{5.9}$$

where $\mu(\mathbf{x}; \boldsymbol{\theta}) = \mathbb{E}_{Y|\mathbf{X}}(y|\mathbf{x})$ and $\mu(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{X}}[\mu(\mathbf{x}; \boldsymbol{\theta})]$. According to equation (5.9), the prevalence of the sampled covariates is proportional to the magnitude of $\dfrac{\mu(\mathbf{x}; \boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}$, which contains information about the parameter $\boldsymbol{\theta}$. This information is, also, reflected in the likelihood function as equations (5.8) and (5.9) imply that

$$\mathcal{L}_{J}(\boldsymbol{\theta}) = \mathcal{L}_{I}(\boldsymbol{\theta}) \prod_{i=1}^{n} \frac{\mu(\mathbf{x}_i; \boldsymbol{\theta}) f_{\mathbf{X}}(\mathbf{x}_i)}{\mu(\boldsymbol{\theta})}, \tag{5.10}$$

which provides the complete relation between the conditional and the joint likelihood functions. Equation (5.10), clearly, shows the distinction between the conditional and joint likelihoods and suggests that, despite the conventional regression approach, when there is no covariate bias involved, $\mathcal{L}_{I}$ and $\mathcal{L}_{J}$ may lead to different outcomes if being employed as basis for maximum-likelihood-based estimation for either parameter estimation or even variable selection. Of course, this distinction is due to the particular sampling procedure we have adopted.

Next, we give the joint likelihood in terms of the unbiased distribution of the covariates as well. Although not very common in practice, this might be of interest in cases where $f_{\mathbf{X}}$ is known. Even if this information is not available, Bergeron [2006] provides a way for estimating $f_{\mathbf{X}}$, empirically from the biased data. Hence, the following form might be useful in either case. Notice that the next formula is the immediate result of replacing $f^{*}_{\mathbf{X}|\overline{T}}$, in equation (5.7), with what equation (5.9) provided:

$$\mathcal{L}_{J}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{ \left[\frac{y_i f_{Y}(y_i|\mathbf{x}_i; \boldsymbol{\theta}) f_{\mathbf{X}}(\mathbf{x}_i)}{\mu(\boldsymbol{\theta})} \right]^{\delta_i} \left[\int_{v \geq a_i + c_i} \frac{v f_{Y}(v|\mathbf{x}_i; \boldsymbol{\theta}) f_{\mathbf{X}}(\mathbf{x}_i)}{\mu(\boldsymbol{\theta})}\,\mathrm{d}v \right]^{1-\delta_i} \right\}.$$

To complete the discussion about the conditional and joint approaches, we will next derive the corresponding log-likelihood functions. The reason is that the maximum likelihood estimates are most often derived by maximizing the log-likelihood function rather than the likelihood itself. This, especially, would come with significant computational advantages, compared to directly using the likelihood function for both purposes, i.e., the estimation of the parameters and selection of the optimal subset of covariates.

However, prior to extracting the log-likelihood functions, we would like to make a "formal" remark. Although this point is of minor importance from the conceptual perspective, when not properly specified, might be quite confusing. In the upcoming paragraphs, we discussion this confusion swiftly.

Oftentimes in statistics and particularly in the MLE context, the objective function to be maximized is neither the likelihood nor the log-likelihood function. Since only parameter-dependent terms of the (log-)likelihood function are determinant of the final estimates of the parameters one, usually, gets rid of the parameter-free terms. Nonetheless, authors usually do not hesitate to refer to these modified versions of the (log-)likelihood functions as "log-likelihood" or "likelihood" functions, without clearly clarifying that what is utilized as the objective function for maximization is not, actually, the (log-)likelihood but a modified (log-)likelihood function that is, optimization-wise, equivalent to the complete likelihood or log-likelihood functions.In the next paragraph, which is devoted to the extraction of the log-likelihood functions, we will try to avoid this confusion.

To extract the conditional and joint log-likelihood functions, firstly, notice that both $\mathcal{L}_{I}$ and $\mathcal{L}_{J}$ have parameter-free terms that might be eliminated. The resulted functions do *not* equal $\mathcal{L}_{I}$ and $\mathcal{L}_{J}$, but are equivalent to them in terms of maximization over the parameter $\boldsymbol{\theta}$. Hence, they will be denoted by $\mathcal{L}_{I}^{*}$ and $\mathcal{L}_{J}^{*}$. That is,

$$\mathcal{L}_{I,n}^{*}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{ \left[\frac{f_{Y}(y_i|\mathbf{x}_i;\boldsymbol{\theta})}{\mu(\mathbf{x}_i;\boldsymbol{\theta})} \right]^{\delta_i} \left[\int\limits_{v \geq a_i + c_i} \frac{f_{Y}(v|\mathbf{x}_i;\boldsymbol{\theta})}{\mu(\mathbf{x}_i;\boldsymbol{\theta})} \, \mathrm{d}v \right]^{1-\delta_i} \right\}, \tag{5.11}$$

and

$$\mathcal{L}_{J,n}^{*}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{ \left[\frac{f_{Y}(y_i|\mathbf{x}_i;\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})} \right]^{\delta_i} \left[\int\limits_{v \geq a_i + c_i} \frac{f_{Y}(v|\mathbf{x}_i;\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})} \, \mathrm{d}v \right]^{1-\delta_i} \right\}. \tag{5.12}$$

Accordingly, define $\ell^*_{I,n}(\boldsymbol{\theta}) := \ln[\mathcal{L}^*_{I,n}(\boldsymbol{\theta})]$ and $\ell^*_{J,n}(\boldsymbol{\theta}) := \ln[\mathcal{L}^*_{J,n}(\boldsymbol{\theta})]$, i.e.,

$$\ell^*_{I,n}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \Big\{ \delta_i \ln f_{Y}(y_i|\mathbf{x}_i;\boldsymbol{\theta}) + (1-\delta_i) \ln \Big[\int\limits_{v \geq a_i + c_i} f_{Y}(v|\mathbf{x}_i;\boldsymbol{\theta})\, \mathrm{d}v \Big] - \ln \mu(\mathbf{x}_i;\boldsymbol{\theta}) \Big\}, \quad (5.13)$$

and

$$\ell^*_{J,n}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \Big\{ \delta_i \ln f_{Y}(y_i|\mathbf{x}_i;\boldsymbol{\theta}) + (1-\delta_i) \ln \Big[\int\limits_{v \geq a_i + c_i} f_{Y}(v|\mathbf{x}_i;\boldsymbol{\theta})\, \mathrm{d}v \Big] \Big\} - n \ln \mu(\boldsymbol{\theta}). \quad (5.14)$$

Obviously, as mentioned before, the following hold:

$$\arg\max_{\boldsymbol{\theta}} \ell^*_I(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \mathcal{L}^*_I(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \mathcal{L}_I(\boldsymbol{\theta}),$$
$$\arg\max_{\boldsymbol{\theta}} \ell^*_J(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \mathcal{L}^*_J(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \mathcal{L}_J(\boldsymbol{\theta}).$$

As one may see in equations (5.11) and (5.12), $\mu(\mathbf{x}_i;\boldsymbol{\theta})$ in $\mathcal{L}^*_{I,n}$ is replaced with $\mu(\boldsymbol{\theta})$ in $\mathcal{L}^*_{J,n}$. Consequently, $\ell^*_{I,n}$ and $\ell^*_{J,n}$ differ from each other in their last terms, i.e., $-\sum_{i=1}^{n} \ln \mu(\mathbf{x}_i;\boldsymbol{\theta})$ is changed to $-n \ln \mu(\boldsymbol{\theta})$, as equations (5.13) and (5.14) depict, and apart from this difference, the rest is the same. For ease, let $\ell^*_{\cap,n}$ denote the common part, i.e.,

$$\ell^*_{\cap,n}(\boldsymbol{\theta}) := \delta_i \ln f_{Y}(y_i|\mathbf{x}_i;\boldsymbol{\theta}) + (1-\delta_i) \ln \Big[\int\limits_{v \geq a_i + c_i} f_{Y}(v|\mathbf{x}_i;\boldsymbol{\theta})\, \mathrm{d}v \Big],$$

and then the simplified log-likelihoods can be expressed as

$$\ell^*_{I,n}(\boldsymbol{\theta}) = \ell^*_{\cap,n} - \sum_{i=1}^{n} \ln \mu(\mathbf{x}_i;\boldsymbol{\theta})$$

and

$$\ell^*_{J,n}(\boldsymbol{\theta}) = \ell^*_{\cap,n} - n \ln \mu(\boldsymbol{\theta}). \quad (5.15)$$

The difference between the last terms is, in fact, the responsible factor for the efficiency gain in parameter estimation when the joint likelihood is applied in lieu of the conditional one.

Now, let us introduce, very briefly, a few characteristics of $\mathcal{L}_I$ and $\mathcal{L}_J$, particularly, in connection with parameter estimation. Although our interest lies in variable selection rather than parameter estimation, a short discussion of these characteristics would be insightful to understand the related aspects of the variable selection problem. This results are available in Bergeron [2006].

5.4.2 Conditional and Unconditional Estimation vs Selection

Now, returning back to the comparison of the two likelihoods, let $\ell_{I,n}$ and $\ell_{J,n}$ denote the corresponding log-likelihood functions calculated through a sample of size n. Similarly, suppose that $\hat{\boldsymbol{\theta}}_{I,n}$ and $\hat{\boldsymbol{\theta}}_{J,n}$ represent the parameter estimates under the conditional and joint likelihoods, respectively.

Bergeron et al. [2008] discussed that $\mathcal{L}_I$ results in *efficiency loss* in the estimation of $\boldsymbol{\theta}$, when applied instead of $\mathcal{L}_J$. Section 3.2 of Bergeron et al. [2008] provides an analytical example where $\hat{\boldsymbol{\theta}}_{J,n}$ is 50% more efficient than its counterpart $\hat{\boldsymbol{\theta}}_{I,n}$. However, it was established that, asymptotically, they provide identical estimates if identifiability is assumed. In particular, it was demonstrated that

$$\frac{1}{n}\left|\frac{\partial}{\partial\boldsymbol{\theta}}\ell_{I,n}(\boldsymbol{\theta})-\frac{\partial}{\partial\boldsymbol{\theta}}\ell_{J,n}(\boldsymbol{\theta})\right|\xrightarrow{a.s.}0,\qquad\text{as } n\to\infty,$$

that is, $\hat{\boldsymbol{\theta}}_{I,n}$ and $\hat{\boldsymbol{\theta}}_{J,n}$ are asymptotically equivalent. Also, it was proved that $\hat{\boldsymbol{\theta}}_{J,n}$ is an asymptotic consistent estimator of $\boldsymbol{\theta}$, i.e.,

$$\hat{\boldsymbol{\theta}}_{J,n}\xrightarrow{a.s.}\boldsymbol{\theta},\qquad\text{as } n\to\infty.$$

Obviously, consistency of $\hat{\boldsymbol{\theta}}_{J,n}$, together with asymptotic equivalence of the two likelihoods, guarantee that $\hat{\boldsymbol{\theta}}_{I,n}$, also, estimates $\boldsymbol{\theta}$, consistently.

To summarize, while both likelihoods provide asymptotically consistent estimators, the efficiency is still a big concern since the estimation efficiency loss might be even more destructive when only a small sample is available. And as it was mentioned earlier, this is a typical scenario in many practical situations. In addition, an important point to note here is that consistency in parameter estimation does not necessarily leads to variable selection consistency. For a variable selection criterion to be consistent, the penalty term, mentioned earlier, plays a crucial role. We will elaborate on this point later in this chapter.

5.4.3 Estimation of the Joint Likelihood

Equation (5.15) reveals an important point: In order to estimate $\ell^*_{J,n}(\boldsymbol{\theta})$, one needs to estimate the overall mean $\mu(\boldsymbol{\theta})$, whose value cannot be directly extracted from prevalent subjects as due to the length bias the overall mean in the prevalent population is biased upward. As a result, In the next part, we focus on the estimation of $\mu(\boldsymbol{\theta})$ from a sample of LBRC-C data. Theoretically, there exist two approaches in order to estimate the incident overall mean; albeit they impose different restrictions, practically speaking. One approach makes

use of the distribution of the *unbiased* covariates, while the other one depends on the biased distribution of the covariates.

First Approach

The incident overall mean $\mu(\boldsymbol{\theta})$ was defined to be the expected value of the conditional mean lifetimes under the unbiased distribution of the covariates, i.e.,

$$\begin{aligned}\mu(\boldsymbol{\theta}) &:= \mathbb{E}_{\mathbf{X}}[\mu(\mathbf{x};\boldsymbol{\theta})] \\ &= \int \mu(\mathbf{u};\boldsymbol{\theta}) f_{\mathbf{X}}(\mathbf{u})\,\mathrm{d}\mathbf{u}.\end{aligned}$$

If the *unbiased* covariate distribution $f_{\mathbf{X}}$ is available, $\mu(\boldsymbol{\theta})$ can be estimated directly from the definition. Unfortunately, in many practical scenarios, having this information is just too unrealistic. On the other hand, we have already mentioned that $f_{\mathbf{X}}$ can be estimated from the biased data provided by Bergeron [2006]. Despite this possibility, we would take another approach because estimating $f_{\mathbf{X}}$, according to Bergeron's method, itself adds a considerable amount of computation, and consequently, decreases the computational efficiency.

Second Approach

On the contrary, the second approach involves using the factorization of $f^{*}_{\mathbf{X}}$, obtained in equation (5.9). This way, one can estimate $\mu(\boldsymbol{\theta})$ without having to, additionally, estimate the distribution of the unbiased covariates. Notice that equation (5.9) can be rearranged as

$$\frac{f_{\mathbf{X}}(\mathbf{x})}{\mu(\boldsymbol{\theta})} = \frac{f^{*}_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta})}{\mu(\mathbf{x};\boldsymbol{\theta})}, \qquad \forall \mathbf{x}.$$

Integrating both sides yields

$$[\mu(\boldsymbol{\theta})]^{-1} = \int \frac{f^{*}_{\mathbf{X}}(\mathbf{u};\boldsymbol{\theta})}{\mu(\mathbf{u};\boldsymbol{\theta})}\,\mathrm{d}\mathbf{u} = \mathbb{E}^{*}_{\mathbf{X}}[\mu(\mathbf{x};\boldsymbol{\theta})^{-1}], \qquad \forall \mathbf{x},$$

which, in turn, suggests the following estimator for the overall mean:

$$\hat{\mu}(\boldsymbol{\theta}) := \left[\frac{1}{n}\sum_{i=1}^{n} \mu(\mathbf{x}_i;\boldsymbol{\theta})^{-1}\right]^{-1}.$$

The last equation gives rise to an estimator of $\ell^*_{J,n}$ defined by

$$\widehat{\ell}^*_{J,n}(\boldsymbol{\theta}) := \ell^*_{\cap,n}(\boldsymbol{\theta}) - n \ln \hat{\mu}(\boldsymbol{\theta}).$$

In other words, what we are able to do in practice is to estimate the joint log-likelihood, in contrast to the typical setting of the MLE. However, in the parametric approach context, the conditional mean $\mu(\mathbf{x}; \boldsymbol{\theta})$ is known and does not need to be estimated, which in turn, makes it possible to calculate the conditional log-likelihood without needing to estimate it.

Now, what is left is to investigate how and why applying the unconditional approach in the context of variable selection provides better results in terms of selecting the optimal subset of the available variables.

5.5 Conditional and Unconditional Variable Selection

Albeit the simulation study in section 6 is based on using model selection criteria for variable selection, we believe that the real effect of replacing $\mathcal{L}_I$ with $\mathcal{L}_J$ should be more visible in the setting where actual variable selection criteria are used. Next, we discuss this point in more detail.

With no loss of generality, assume that the free parameter $\boldsymbol{\theta}$ consists of the regression coefficients only, i.e., $\boldsymbol{\theta} = \boldsymbol{\beta}$. This is a reasonable assumption since our goal is to study the effects of using the joint likelihood in variable selection, compared to the conditional one. Before everything else, let us introduce the variable selection criteria we will consider here. The following two criteria are constructed based on the conditional and unconditional approaches, respectively, and are supposed to be applied to a set of independent LBRC-C training examples:

$$\mathcal{C}_I(\boldsymbol{\beta}) := \ell^*_{I,n}(\boldsymbol{\beta}) - \boldsymbol{P}(\eta, \boldsymbol{\beta}), \tag{5.16}$$

and

$$\mathcal{C}_J(\boldsymbol{\beta}) := \widehat{\ell}^*_{J,n}(\boldsymbol{\beta}) - \boldsymbol{P}(\eta, \boldsymbol{\beta}), \tag{5.17}$$

with $\boldsymbol{P}(\eta, \boldsymbol{\beta})$ being a penalty term defined by

$$\boldsymbol{P}(\eta, \boldsymbol{\beta}) := \sum_{j=1}^{d} \mathcal{P}_\eta(|\beta_j|)$$

where $\mathcal{P}_\eta$ is a suitable penalizing function, possibly depending on n, and $\eta > 0$ is a tuning parameter. The penalty term $\boldsymbol{P}(\eta, \boldsymbol{\beta})$ must satisfy some certain conditions in order for the selection criteria to work properly. Here, we assume that these conditions hold and skip

discussing the details of necessary conditions a penalty term must satisfy as these technical details are out of our interest in this note. One may refer to Antoniadis and Fan [2001] for a detailed discussion of these criteria.

In order to select variables, one maximizes the criteria above over the parameter space $\mathcal{B}$. Here, we assume that the value chosen for the tuning parameter η is the same in both criteria (5.16) and (5.17). The hypothesis of interest is that, with the same tuning parameter η, criterion $\mathcal{C}_{I}$ is more probable to select an incorrect or underfitted model, in comparison with $\mathcal{C}_{J}$. In the sequel, we introduce and discuss a fact that we believe is, probably, one of the key factors which may explain the correctness of the hypothesis. Note that what follows does not prove the validity of the aforementioned hypothesis. Providing a rigorous mathematical proof requires further investigation. Nevertheless, the simulation study provided in the next chapter seems to be aligned with our hypothesis.

We state the following lemma for an arbitrary vector of free parameters, i.e., $\boldsymbol{\theta}$ might include not only the regression coefficients.

Proposition 1. *Let $\ell^{*}_{I,n}(\boldsymbol{\theta})$ and $\widehat{\ell}^{*}_{J,n}(\boldsymbol{\theta})$ be as defined earlier. Then, for any given set of data $\mathcal{D}$ and for any parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, holds the following:*

$$\ell^{*}_{I,n}(\boldsymbol{\theta}) \leq \widehat{\ell}^{*}_{J,n}(\boldsymbol{\theta}), \tag{5.18}$$

with equality happening if and only if $\mu(\mathbf{x}_{1};\boldsymbol{\theta}) = \mu(\mathbf{x}_{2};\boldsymbol{\theta}) = \cdots = \mu(\mathbf{x}_{n};\boldsymbol{\theta})$, i.e., if all the conditional means in the sample are the same.

Proof. To see that, denote $M = \{\mu(\mathbf{x}_{i};\boldsymbol{\theta}) : i = 1,\ldots,n\}$, i.e., the set of all conditional lifetimes corresponding to each subject in sample data. Let G, and H be the geometric and the harmonic mean of M, respectively:

$$G = \left[\prod_{i=1}^{n} \mu(\mathbf{x}_{i};\boldsymbol{\theta})\right]^{\frac{1}{n}}, \quad \text{and} \quad H = \left[\frac{1}{n}\sum_{i=1}^{n} \mu(\mathbf{x}_{i};\boldsymbol{\theta})^{-1}\right]^{-1}.$$

Recall that in our setting both G and H are non-negative. Both $\ell^{*}_{I,n}$ and $\widehat{\ell}^{*}_{J,n}$ can be rephrased easily by means of G and H as follows:

$$\begin{aligned}\ell^{*}_{I,n}(\boldsymbol{\theta}) &= \ell^{*}_{\cap,n}(\boldsymbol{\theta}) - \sum_{i=1}^{n} \ln \mu(\mathbf{x}_{i};\boldsymbol{\theta}) \\ &= \ell^{*}_{\cap,n}(\boldsymbol{\theta}) - \ln G^{n},\end{aligned}$$

and

$$\begin{aligned}\widehat{\ell}^{*}_{J,n}(\boldsymbol{\theta}) &= \ell^{*}_{\cap,n}(\boldsymbol{\theta}) - n \ln \hat{\mu}(\boldsymbol{\theta}) \\ &= \ell^{*}_{\cap,n}(\boldsymbol{\theta}) - \ln H^{n}.\end{aligned}$$

This two equations, together with the well-known fact that $0 \leq H \leq G$, imply inequality (5.18), immediately. □

Now, let us briefly discuss the impact of the employment of $\mathcal{C}_{J}$ instead of $\mathcal{C}_{I}$. Since, for any $\boldsymbol{\beta} \in \mathcal{B}$, we have that $\ell^{*}_{I,n}(\boldsymbol{\beta}) \leq \widehat{\ell}^{*}_{J,n}(\boldsymbol{\beta})$, when one maximizes these criteria, the penalty term $\boldsymbol{P}(\eta, \boldsymbol{\beta})$ might appear to impose heavier penalties to compensate for smaller $\ell^{*}_{I,n}(\boldsymbol{\beta})$ in $\mathcal{C}_{I}(\boldsymbol{\beta})$ compared to $\widehat{\ell}^{*}_{J,n}(\boldsymbol{\beta})$ in $\mathcal{C}_{J}(\boldsymbol{\beta})$. This heavier penalty may result in stronger shrinkage of the coefficient $\boldsymbol{\beta}$ in the conditional criterion $\mathcal{C}_{I}$. Unfortunately, this is not enough to deduce that $\mathcal{C}_{I}$ is more probable to zero out coefficients in comparison to its joint counterpart.

While Proposition 1 point-wise information on the distinction between $\ell^{*}_{I,n}$ and $\widehat{\ell}^{*}_{J,n}$, i.e., at each $\boldsymbol{\beta} \in \mathcal{B}$, what also needs to be investigated is possible geometrical differences between the two functions form the global perspective. This question as well as some other important questions regarding the effects of the joint approach on variable selection, in comparison to the conditional one, are subject to further research in future.

CHAPTER 6 A BRIEF SIMULATION STUDY

Now, let us consider selecting variable by applying model selection criteria, such as the AIC and the BIC. Note that the AIC and BIC are essentially *model selection* criteria, but clearly could be used for selecting variables too. All in all, variable selection is a special case of model selection.

As we have already mentioned, in practice, using the likes of these two criteria for variable selection in high-dimensional data is too wasteful. Nonetheless, it would be insightful to study the distinctions between the behaviours of the considered conditional and joint likelihoods in the context of variable selection by the AIC or BIC for the following reason: Studying the BIC, for example, makes it possible to follow changes in the BIC values as different combinations of variables are included in the model. This is not possible if one uses the criteria designed specifically for variable selection like the LASSO or SCAD. The reason is that, with these criteria, the selection and estimation take place simultaneously, as explained earlier in this chapter.

In the simulation study, provided in subsection 6, we focus on the BIC for the reasons discussed above.

In this numerical example, we illustrate how the choice of the likelihood function, i.e., conditional or joint, would affect the output of the variable selection by the BIC. Further, let BIC_I and BIC_J represent the BIC based on the conditional and joint likelihood, respectively.

6.1 Description of the Incident Population

Here, we assume that the failure time in the incident population follows an exponential distribution, i.e., $Y \sim \text{Exp}(\lambda)$, $\lambda > 0$. In presence of covariates, instead of assuming a single exponential distribution for the entire population, we deal with a mixture of exponential failure times. In other words, given a covariate vector $\mathbf{X}$, an exponential distribution $Y|\mathbf{X} \sim \text{Exp}(\lambda e^{\mathbf{X}\boldsymbol{\beta}})$ is assumed to underlie the failure times; in other words, the parameter of the distribution is dependent on the covariates.

To bold the effect of the likelihood on variable selection, we assume that there is no censoring involved, since the difference in likelihoods is due to the length bias.

Suppose that the true model contains a two-dimensional covariate, i.e., attached to each failure time Y_i there is a vector $\mathbf{X}_i = (X_{i_1}, X_{i_2})$ of covariates. Since there is no censoring in-

volved, each datum might be considered of the form $(Y_i, \mathbf{X}_i)$. Further, let $X_{i_j} \sim \text{Bernoulli}(\frac{1}{2})$, $j = 1, 2$, and $\boldsymbol{\beta}$ be the vector of the regression coefficients.

6.2 Derivation of the Likelihoods

We derive $\ell^*_{I,n}$ and $\ell^*_{J,n}$, whose general form was obtained in equations (5.13) and (5.14) for the case of exponentially distributed data as described in the previous section. Recall that $\boldsymbol{\theta}$ denotes the vector of all parameters, regardless of being known or unknown. Thus, in the case of exponential distribution $\boldsymbol{\theta} = (\boldsymbol{\beta}, \lambda)$.

Firstly, let us compute the overall mean $\mu(\boldsymbol{\theta})$. Since $\mu(\mathbf{X}_i; \boldsymbol{\theta}) = (\lambda e^{\mathbf{X}_i\boldsymbol{\beta}})^{-1}$, we have that

$$\mu(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{X}_i}[\mu(\mathbf{x}_i; \boldsymbol{\theta})] = \frac{\lambda^{-1}}{2^d} \sum_{\mathbf{z}\in\{0,1\}^d} e^{-\mathbf{x}\boldsymbol{\beta}}, \tag{6.1}$$

where $d = \dim(\mathbf{X}_i)$. For the sake of simulation we assumed that $\mathbf{X}_i$ is of dimension two, and hence, plugging $d = 2$ in equation (6.1) gives the overall survival time:

$$\mu(\boldsymbol{\theta}) = \frac{1}{4\lambda}\left(1 + e^{-\beta_1} + e^{-\beta_2} + e^{-\beta_1-\beta_2}\right).$$

Following equation (3.13), the density of the length-biased survival time is given by

$$f_{Y_i}(y_i|\mathbf{X}^*_i = \mathbf{x}_i; \boldsymbol{\theta}) = (\lambda e^{\mathbf{x}_i\boldsymbol{\beta}})^2 y_i \exp(-e^{\mathbf{x}_i\boldsymbol{\beta}}\lambda y_i).$$

Notice that the right-hand side of the last equation equals the density of the Gamma distribution. It is well known that if $Y_i|\mathbf{X}_i \sim \text{Exp}(\lambda e^{\mathbf{X}_i\boldsymbol{\beta}})$, then the length-biased sampling distribution of the lifetimes conditioned on the biased covariate $\mathbf{X}^*_i$ follows the Gamma distribution with shape parameter 2 and rate $\lambda e^{\mathbf{X}^*_i\boldsymbol{\beta}}$, i.e., $\tilde{Y}_i|\mathbf{X}^*_i \sim \text{Gamma}(2, \lambda e^{\mathbf{X}^*_i\boldsymbol{\beta}})$.

Now, having the density, as well as the conditional and overall means $\mu(\mathbf{x}_i; \boldsymbol{\theta})$ and $\mu(\boldsymbol{\theta})$, equation (5.9) gives the biased sampling distribution of the covariates as follows:

$$\mathbb{P}(\mathbf{X}^*_i = \mathbf{x}_i) = \frac{e^{-\mathbf{x}_i\boldsymbol{\beta}}}{1 + e^{-\beta_1} + e^{-\beta_2} + e^{-\beta_1-\beta_2}}, \quad \mathbf{x}_i \in \{0, 1\}^2.$$

Finally, given some training data, consisting of n independently distributed observations, $\ell^*_{I,n}$ and $\ell^*_{J,n}$ can be computed from equations (5.13) and (5.14) as follows:

$$\ell^*_{I,n}(\boldsymbol{\theta}) = 2n \ln \lambda + \sum_{i=1}^{n} (2\mathbf{x}_i \boldsymbol{\beta} - e^{\mathbf{x}_i \boldsymbol{\beta}} \lambda y_i),$$

$$\ell^*_{J,n}(\boldsymbol{\theta}) = 2n(\ln 2 + \ln \lambda) + \sum_{i=1}^{n} (\mathbf{x}_i \boldsymbol{\beta} - e^{\mathbf{x}_i \boldsymbol{\beta}} \lambda y_i) - \\ - \sum_{i=1}^{n} \ln(1 + e^{-\beta_1} + e^{-\beta_2} + e^{-\beta_1 - \beta_2}).$$

6.3 Data Simulation Steps

In order to generate one dataset of size n we perform the following steps in the given order:

1. *Generation of the Incident Population*

   (a) Let N is the incident population size. A large enough size must be chosen. First, generate N covariates $\mathbf{X}_i = (X_{i_1}, X_{i_2}),\ i = 1, 2, \ldots, N$, such that $X_{i_j} \sim$ Bernoulli($\frac{1}{2}$), for $j = 1, 2$.

   (b) Next, fix a value $\boldsymbol{\beta}^0$ as the true regression coefficient, as well as λ the parameter of the exponential distribution.

   (c) Produce lifetimes $Y_i,\ i = 1, 2, \ldots, N$ according to $Y_i | \mathbf{X}_i \sim \text{Exp}(\lambda e^{\mathbf{X}_i \boldsymbol{\beta}^0})$.

   (d) Finally, attach the covariates to their corresponding lifetimes to form the incident population $(Y_i, \mathbf{X}_i),\ i = 1, 2, \ldots, N$.

2. *Generation of the Truncation Times*

   (a) First, fix an appropriate value T^0 as the upper bound for the truncation time.

   (b) As mentioned before, stationarity requires the truncation times to be uniformly distributed. Hence, simulate a truncation time A_i for each observation in the population such that, $A_i \sim \text{Unif}(0, A^0),\ i = 1, 2, \ldots, N$.

   (c) Lastly, complete the generation of the incident population by adding the truncation times to their corresponding pairs obtained in (1-d) above. Now, each datum is of the form $(Y_i, \mathbf{X}_i, A_i)$.

3. *Prevalent Population Extraction*

(a) Detect and separate the subjects for which holds $Y_i > A_i$. These subjects comprise the prevalent population. The rest of the subjects are said to be left truncated and will not be considered for sampling.

4. *Recruiting Subjects (Sampling)*

   (a) Randomly, select n subjects from the prevalent cohort. Now, the selected samples have the desired distribution and, hence, of the form $(\tilde{Y}_i, \overset{*}{\mathbf{X}}_i, \tilde{A}_i)$. See Figure 6.1 for and example of a simulated dataset.

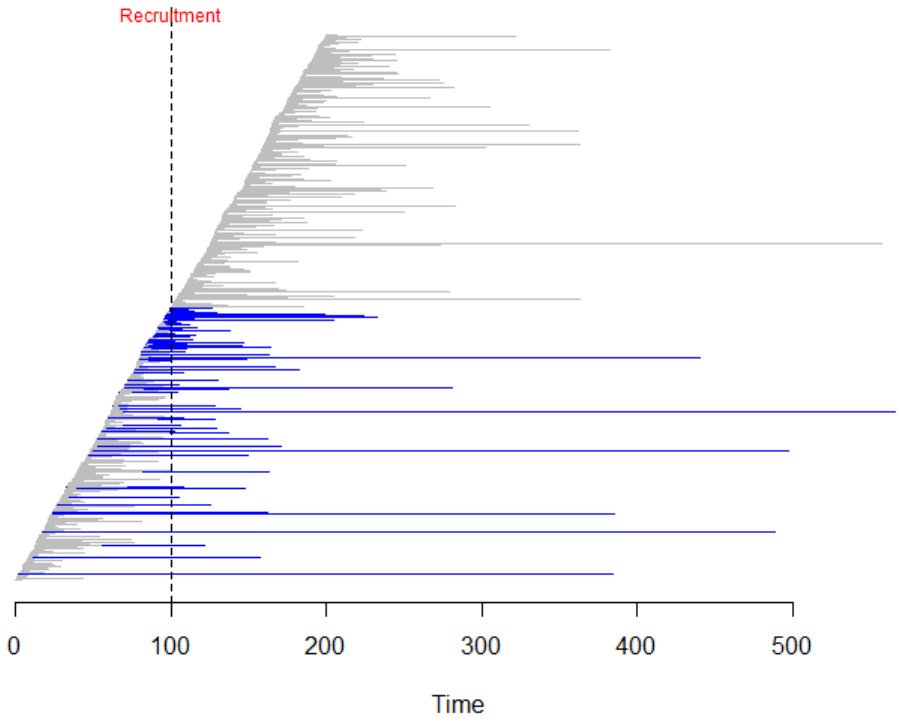

Figure 6.1 **Simulated Incident and Prevalent Populations.** An example of a simulated population according to the steps explained in this section is depicted. The horizontal line represents time in *days*. Zero is, in fact, the moment when the first onset in the incident population appeared. Day 100 is the recruitment time, showed by the vertical dashed line. Blue horizontal lines are subjects in the prevalent population, while their union with light grey lines makes the incident population. For better visualization, subjects have been sorted by the onset and with a slight vertical distance between each pair.

The procedure above, however, generates one sample set. For illustrating the probability of choosing an incorrect model, we repeat these steps multiple times, for each sample size n, to generate enough data. In our simulation, we have used 39 sample sizes, gradually growing from small amounts to larger ones. For each individual sample size, 50 different datasets have been generated. More numerical details are provided in later sections.

6.4 Candidate Models

Denote the set of candidate models by $\mathcal{M}$. Then, we have that $\mathcal{M} = \{M_1, M_2, M_{1,2}\}$, where

$$M_1 = \Big\{\boldsymbol{\beta} \in \mathcal{B} \mid \mathcal{N}(\boldsymbol{\beta}) = \{1\}\Big\},$$
$$M_2 = \Big\{\boldsymbol{\beta} \in \mathcal{B} \mid \mathcal{N}(\boldsymbol{\beta}) = \{2\}\Big\},$$
$$M_{1,2} = \Big\{\boldsymbol{\beta} \in \mathcal{B} \mid \mathcal{N}(\boldsymbol{\beta}) = \{1,2\}\Big\}.$$

We choose the true regression coefficients to belong to $M_{1,2}$ to rule out the possibility of overfitting by either criteria BIC_I or BIC_J. This way we give more chances to both criteria to underfit as selecting an incorrect model is what we would like to measure and compare between the two.

6.5 Numerical Results

Here, we present a summary of the numerical results obtained from variable selection using the Bayesian information criterion (BIC), once based on the conditional likelihood $\ell^*_{I,n}$ and then by the joint likelihood $\widehat{\ell}^*_{J,n}$. The simulation was implemented in the programming language `R`.

In simulation, 39 sample sizes $n = 50, 75, \ldots, 1000$, were used. For each n we generated 50 different length-biased datasets according to the procedure demonstrated earlier in this section. In order to perform selection, the following steps were performed:

for each sample size $n = 50, 75, \ldots, 1000$:

1. for each sample dataset $\mathcal{D}_{n,r}$, $r = 1, 2, \ldots, 50$:
   (a) for each model in $\mathcal{M}$:
      i. estimate coefficients
      ii. compute BIC_I
      iii. compute BIC_J
   (b) extract the optimal model by BIC_I
   (c) extract the optimal model by BIC_J
2. calculate $p_{I,n}$ (percentage of incorrect models by BIC_I)
3. calculate $p_{J,n}$ (percentage of incorrect models by BIC_J)

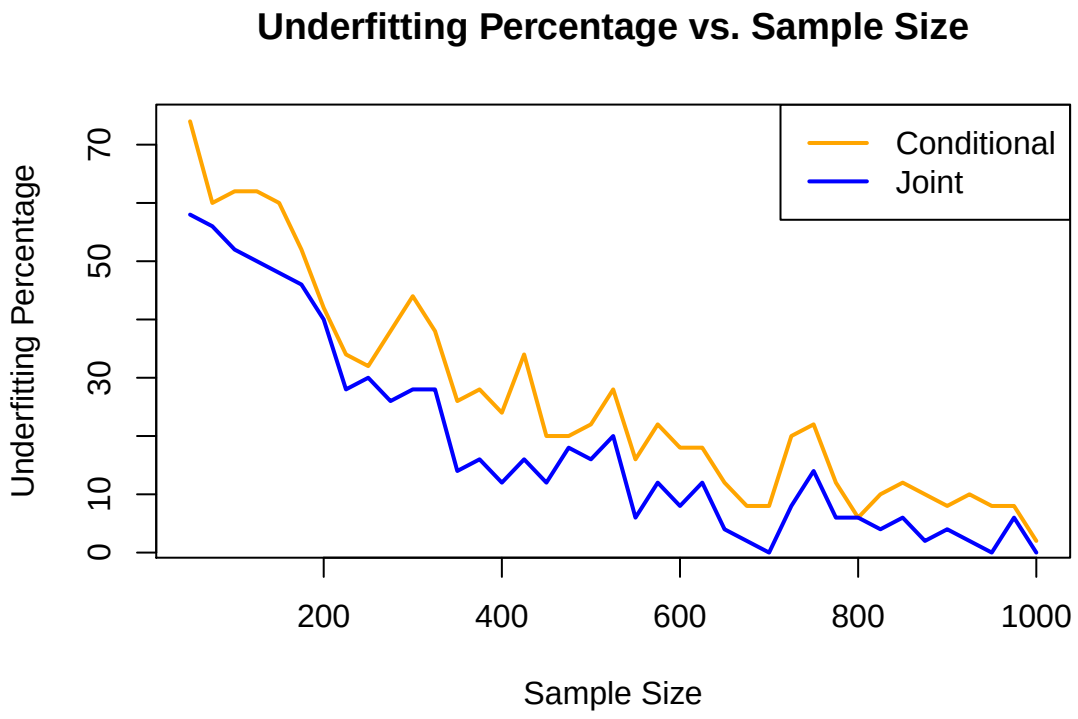

Figure 6.2 **Incorrect Selections' Percentages by the Conditional and Joint Approaches.**

As sample size n gradually increases, the underfitting percentage for both conditional and joint likelihoods approaches 0. Clearly, regardless of the value of n, variable selection using the BIC_J results in a less percentage of underfitting (the blue line) compared to that of the BIC_I (the orange line). Also, when n is smaller, the difference between the percentages is slightly bigger.

Table 6.1 Incorrect Selections' Percentages by the Conditional and Joint Approaches

n	$p_{I,n}$	$p_{J,n}$
50	74	58
75	60	56
100	62	52
125	62	50
150	60	48
175	52	46
200	42	40
225	34	28
250	32	30
275	38	26
300	44	28
325	38	28
350	26	14
375	28	16
400	24	12
425	34	16
450	20	12
475	20	18
500	22	16
525	28	20
550	16	06
575	22	12
600	18	08
625	18	12
650	12	04
675	08	02
700	08	0
725	20	08
750	22	14
775	12	06
800	06	06
825	10	04
850	12	06
875	10	02
900	08	04
925	10	02
950	08	0
975	08	06
1000	02	0

CHAPTER 7 A SURVEY OF LEARNING BY NEURAL NETWORKS

In Chapter 3, we have seen that the main problem of statistical learning theory is function estimation from a limited amount of empirical data. Also, in Subsection 1.5.5, we mentioned that there exist numerous function estimation methods, such as the MLE, SVMs, and many others, which may be applied for the purpose of function estimation. The current chapter focuses on one of the existing estimation tools from empirical data, called mapping neural networks. A mapping neural network is a perceptron-type neural network, consisting of multiple connected processing units or nodes, each of which being able to accomplish specific types of simple computation. An appropriately chosen collection of nodes can approximate certain classes of functions at any point of their domains with arbitrary precision. In other words, a mapping neural network might, potentially, provide a concrete tool to solve the main problem of learning. The mapping neural network's capability to solve the learning problem and, especially, their promising performance in several applications has brought them a tremendous amount of attention.

Efficient use of the neural networks, generally, and in the realm of incomplete data, specifically, requires a deep understanding of the fundamental aspects of the networks. To begin with, a rigorous and systematic establishment of their mathematical foundations with respect to learning from complete, i.i.d. data should be the very first step. The content of this chapter can be viewed as an initial attempt to fulfil this first step. Therefore, firstly, we investigate the mathematical genealogy of the mapping neural networks and, secondly, provide the connection amongst the results that justify the networks' estimation capability. While our eventual objective is to study the mapping neural networks' ability to learn from biased and censored data, we restrict this note to the case of i.i.d. data for the time being and leave this ultimate goal for future studies.

7.1 Function Representation and Related Problems

In 1900, at the International Congress of Mathematicians in Paris, David Hilbert presented 10 unsolved problems that he believed were some of the most fundamental problems of mathematics to be tackled in the 20th century. Later, he published a list of 23 problems, including those 10 introduced at the conference, which are currently known as Hilbert's problems. Some of these problems became part of the most inspirational questions for mathematicians throughout the entire century and beyond. Subsequently, an immense amount of push and power has been placed in order to shed some light on these problems. Luckily, these attempts

paid off and resulted in some influential breakthroughs in other fields of science.

One such problem is the thirteenth problem, which triggered a tremendous amount of interest amongst numerous mathematicians, which in turn, leads to the creation of a huge corpus of works contributing to this problem and related issues, primarily, algebraic theory of functions. We will see that the impacts did not remain limited to the horizons of algebraic theory of functions but rather spread to other fields, such as functional analysis, topology, abstract algebra, and etc. A variety of applications were, also, emerged as a result of the endeavours carried out around the thirteenth problem.

In this section, we study the role of this problem and related issues in the analysis of one of the most popular algorithms of learning, i.e., the *mapping neural networks*. In particular, we investigate how their capability of function estimation might be justified from the theoretical point of view. We also explore different attempts that have been made to either prove or disprove Hilbert's thirteenth problem.

Neural networks began to prosper mainly during the last decades of the previous century. Nevertheless, one may trace their mathematical roots back to the very beginning of the 20th century and, in some cases, even earlier. There are a series of mathematical findings that are related to the computational structure of the mapping neural networks. These are valuable facts, which can explain some less investigated aspects of the problem regarding the ability of the mapping neural networks to solve the learning main problem. Here, we discuss the relations among these events and try to connect the most important ones in a way that facilitates shaping the skeleton of the neural networks.

To this end, we sometimes flash back to earlier periods of the development of mathematics, which we suppose may help draw a more accurate picture of the flow of events in its entirety. Throughout the chapter, the necessary technical details are introduced as the discussion advances.

7.1.1 Hilbert's Thirteenth Problem: The Original Statement

Hilbert, formulated the 13th problem under the title *"Impossibility of the Solution of the General Equation of the 7th Degree by Means of Functions of Only 2 Arguments"*, according to the English translation of his work [Hilbert, 1902] [1]. Apparently, the problem deals with the solutions of the following equation:

$$a_7x^7 + a_6x^6 + a_5x^5 + a_4x^4 + a_3x^3 + a_2x^2 + a_1x + a_0 = 0.$$

[1] The original work, in German language, was published earlier in *Göttinger Nachrichten* [Hilbert, 1900] and was translated to English by Mary Frances Winston Newson.

The problem of resolvents was not a new problem at that time and had been studied by many mathematicians before Hilbert. However, what made Hilbert's problem distinguished from the preceding ones was perhaps the novel perspective to the possible solution of polynomial equations. While the main interest lied in algebraic *solvability* of polynomial equations before him, the thirteenth problem focuses on the complexity of the solutions in a certain sense. The solvability problem involves answering the question about the *existence* of *algebraic solutions* (also, called *solutions in radicals*): Assume the general polynomial equation of degree n, i.e.,

$$P_n(x)[f(x)]^n + P_{n-1}(x)[f(x)]^{n-1} + \cdots + P_1(x)[f(x)] + P_0(x) = 0, \tag{7.1}$$

where $P_n(x) \neq 0$, and $P_0, \ldots, P_n$ are known polynomials in x with integer or, equivalently, rational coefficients, $f(x)$ an unknown function of x, and x itself a complex or real variable. Now, the question of algebraic solvability consists of determining whether the solutions of equation (7.1) can be expressed by closed-form algebraic expressions over the field of rational numbers; that is, by finite expressions in terms of the coefficients $P_0, \ldots, P_n$ and algebraic operations of *addition, subtraction, multiplication, division,* and *rational exponentiation.*

While equations up to degree 4, inclusive, can be solved in radicals, their counterparts of higher degrees, generally, cannot. This was known long before Hilbert; for example, due to the *Abel-Ruffini* theorem, which states that the general polynomials of degree 5 or higher are not solvable by radicals. A similar but stronger result from Galois theory implies that unsolvable polynomials of degree 5 or higher constitute an *everywhere dense* subset of the polynomials of the corresponding degrees.

However, Hilbert's concern was different in that the complexity of the solutions, in the sense of the minimum number of arguments necessary for expressing the solutions, was the central point of the problem. He stated the problem in connection with *nomography* [Hilbert, 1902]. A *nomogram* or *nomograph* is a tool, invented by Philbert Maurice d'Ocagne, for computing functions and their roots by means of simple graphical methods. Especially, nomographic tables could effectively solve the equations whose solutions could be expressed by functions of maximum two variables. Therefore, it was important for the solutions of an equation to be expressible by successive compositions of functions of maximum two variables in order to be solvable by nomographic constructions. This was the motivation behind Hilbert's thirteenth problem.

Existence of algebraic solutions for equations of degree less than or equal to 4 implies that one can reduce their solutions to compositions of bivariate functions. For equations of degree five and higher, one can apply Tschirnhausen transformations, which readily transform the equation to a form being solvable by $(n-4)$-variate functions. Note that the transformation

is carried out algebraically. Therefore, the solutions of the equations of degree 5 and 6 could be immediately represented by functions relying on one and two variables, respectively.

So far, it was showed that all of the equations with degrees strictly less than seven are solvable by functions of maximum two variables. So, the general equation of degree seven is the smallest degree that cannot be solved by functions of maximum two variables, and consequently, by nomographic tables (A Tschirnhausen transformation brings the number of the needed variables down only to three). This was, in fact, the context which motivated Hilbert to postulate his conjecture.

As we will see later, the original statement of the thirteenth problem by Hilbert can be understood differently. This, actually, led to multiple speculations about the problem, each of which has opened a new direction of viewing and tackling the problem. The following formulation of the problem is due to the first English translation of Hilbert's paper, which was originally published in German [Hilbert, 1900].[2] There, Hilbert mentions that it *might be* impossible to rewrite the roots by a finite number of successive substitutions of functions of only two arguments. Immediately, he tries to give a more precise formulation of the problem, however, the main intention of him remains rather vague, particularly, the specific assumptions and conditions he meant to impose over the functions in question. Seemingly, *continuity* of the functions were added in later versions of the problem, whereas the earlier ones mentioned that the functions were to be *algebraic* instead. For more details see Vitushkin [2004], among others. Finally, here is the original statement of the problem by Hilbert:

> *"The general equation of the seventh degree, i.e.,* $f^7 + xf^3 + yf^2 + zf + 1 = 0$, *is not solvable with the help of any continuous functions of only two arguments."*

7.1.2 Functions of Three Variables Do Not Exist

The title of this subsection has been inspired by a question in the famous book *Problems and Theorems in Analysis I* by George Pólya and Gabor Szegö, originally published in Berlin in 1925. Questions 119 and 119*a* in the second part of the third chapter of the book deal with possible representations of functions of 3 variables by means of functions of less than 3 variables [Pólya and Szegö, 1998].

More precisely, let $f : \mathbb{R}^3 \to \mathbb{R}$ be an arbitrary function. Is it always possible to find functions $\varphi : \mathbb{R}^2 \to \mathbb{R}$ and $\psi : \mathbb{R}^2 \to \mathbb{R}$ such that for every $(x, y, z) \in \mathbb{R}^3$, holds $f(x, y, z) = \psi(\varphi(x, y), z)$? What if all the f, φ, and ψ assumed to be continuous?

[2] The translation [Hilbert, 1902], is due to Mary Frances Winston Newson, the first female American mathematician who received a PhD in mathematics from a European university (University of Göttingen, Germany).

Before looking into the details of the question above, let us state one of the essential concepts in our further discussion; the so-called *superposition* of functions:

Definition 13 (Superposition of Functions)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be an arbitrary function and m be a natural such that $1 \leq m < d$. Then, f is said to be a superposition of functions of maximum m variables if f can be written as*

$$f(x_1, \ldots, x_d) = g_0(y_{(0,1)}, y_{(0,2)}, \ldots, y_{(0,k_0)}),$$

where each argument of the function g_0 can be decomposed by the recurrent formula

$$y_{(\beta_0,\ldots,\beta_i)} = g_{(\beta_0,\ldots,\beta_i)}(y_{(\beta_0,\ldots,\beta_i,1)}, y_{(\beta_0,\ldots,\beta_i,2)}, \ldots, y_{(\beta_0,\ldots,\beta_i,k_i)}),$$

where $i = 1, \ldots, s$, with $s \in \mathbb{N}$, $\beta_0 = 0$, $\beta_j = 1, 2, \ldots, k_{j-1}$, with k_{j-1} being the dimension of the argument of $g_{(\beta_0,\ldots,\beta_{j-1})}$. Note that for any $(\beta_0, \ldots, \beta_i)$, we have that $K_i \leq m$.

Example 7.1. Here, we provide three simple functions from $\mathbb{R}^d$ to $\mathbb{R}$, where d equals $3, 3$, and 4, respectively, and give a superposition of each of them using functions of maximum two variables.

1. Let $f(a, b, c) = b$. Define $\varphi_1(x, y) = x$ and $\varphi_2(x, y) = y$. Then,

$$f(a, b, c) = \varphi_1\Big(\varphi_2(a, b), c\Big).$$

2. If $g(a, b, c) = abc$, then define a simple function $\varphi(x, y) = x, y$, and consequently, $g(a, b, c)$ can be written as follows:

$$g(a, b, c) = \varphi\Big(\varphi(a, b), c\Big).$$

3. Let

$$h(a, b, c, d) = \frac{-b + \sqrt{b^2 - 4ac}}{(\sqrt[3]{d} + a)^c}.$$

Define

$$\begin{aligned} &\varphi_1(x, y) = xy, && \varphi_2(x, y) = x + y, && \varphi_3(x) = \sqrt[3]{x}, \\ &\varphi_4(x, y) = -x + \sqrt{x^2 - 4y}, && \varphi_5(x, y) = x^y, && \varphi_6(x) = x^{-1}. \end{aligned}$$

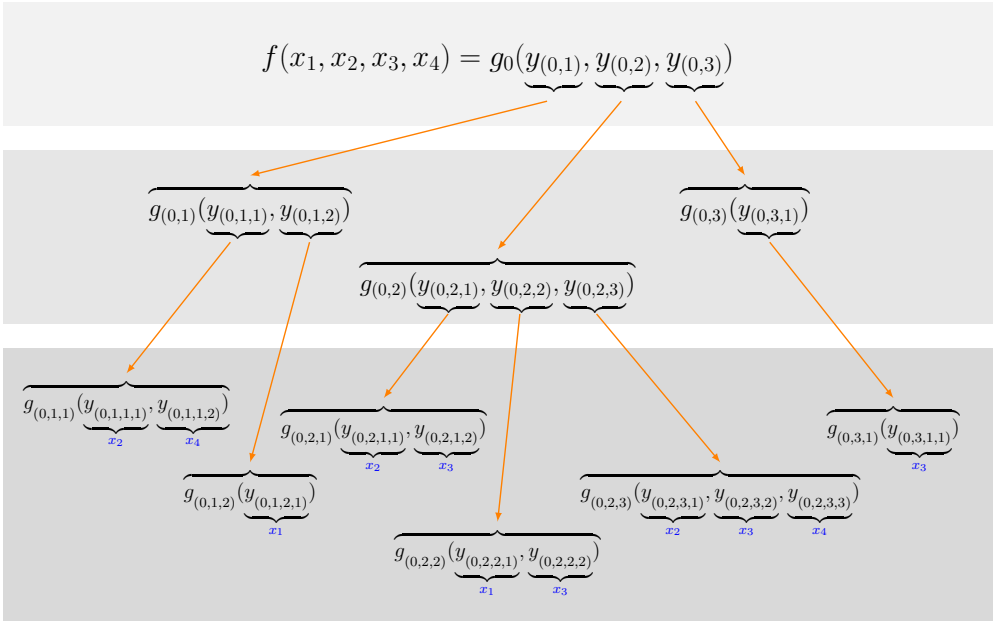

Figure 7.1 **An Example of Superposition of Functions.** Here is an example of superposition of a function f, of 4 variables x_1, x_2, x_3, x_4, by means of functions of maximum 3 variables.

Then, we have that

$$h(a,b,c,d) = \varphi_1\left(\varphi_4\left(b, \varphi_1(a,c)\right), \varphi_6\left(\varphi_5\left(\varphi_2\left(\varphi_3(d), a\right), c\right)\right)\right).$$

Now, returning to the question of the existence of functions of three variables, we start from a simple case, where there is no assumption of continuity. It is not hard to see if discontinued functions are also allowed, then any function of 3 variables can be written in the form of superpositions of functions of 2 variables. More details may be found in Pólya and Szegö [1998], Arnold [2009a].

A surprising point is that most of the earlier results on Hilbert's 13th problem considered somehow more restricted forms of superposition, e.g., allowing only single superpositions. This is the reason why they all approved Hilbert's conjecture. As Pólya and Szegö mention, even a very simple function $f(x,y,z) = xy + yz + zx$ cannot be expressed as a single superposition of bivariate functions. Probably, the most interesting result among this series of results was the one obtained by A. G. Vitushkin, who succeeded to construct a polynomial for which there is a neighbourhood of functions, under the metric induced by the uniform norm, non of which can be decomposed by single superpositions of bivariate continuous functions in any system of coordinates [Vitushkin, 1955].

There also can be other restrictions imposed on the form of functions employed in superposi-

tions. For example, a closely related question to the one appeared in Pólya and Szegö's book could be obtained by replacing the *continuity* condition with being *analytic*. This question has already been answered by Hilbert himself:

> *Not all analytic functions of* 3 *variables can be represented as superpositions of analytic functions of two variables.*

When the space of interest consists of continuous functions, Kolmogorov proved, in 1956, that *every continuous function defined on the d-dimensional unit interval* $I^d = [0,1]^d$*, with* $d \geq 4$*, is a superposition of continuous functions of* 3 *variables*. More precisely, he showed that for any $d \geq 3$, there exist real-valued, continuous functions ϕ_i, ψ_i, $i = 1, 2, \ldots, d$, defined on I^{d-1}, such that any continuous, real-valued function f, defined on I^d, can be represented as follows:

$$f(x_1, x_2, \ldots, x_d) = \sum_{i=1}^{d} h_i\Big[\phi_i(x_1, \ldots, x_{d-1}), \psi_i(x_1, \ldots, x_{d-1}), x_d\Big], \tag{7.2}$$

where h_is are continuous functions defined on I^3 [Kolmogorov, 1961, Theorem 3].[3] The details of the proof provided by Kolmogorov are beyond the scope of our discussion, however, it is worth noting that it is, particularly, interesting because of its topological nature. (Also, see Arnold [2009b]).

Shortly after Kolmogorov proved the aforementioned result, i.e., the possibility of reducing any continuous function of 3 or more variables to a superposition of continuous functions of 3 variables, his student, V. I. Arnold improved it slightly by proving that for the function f, defined in equation (7.2), we have that

$$f(x_1, x_2, \ldots, x_d) = \sum_{i=1}^{d} \Big\{\phi_i[u_i(x_1, \ldots, x_{d-1}), x_d] + \psi_i[v_i(x_1, \ldots, x_{d-1}), x_d] + \chi_i[w_i(x_1, \ldots, x_{d-1}), x_d]\Big\}, \tag{7.3}$$

where all functions ϕ_i, ψ_i, χ_i, and u_i, v_i, w_i are continuous and real-valued for $i = 1, 2, \ldots, d$ [Arnold, 2009b].

The latter result, actually, answers the question posed by Pólya and Szegö, regarding the "existence" of functions of three variables. In order to see that, it is enough to rewrite equation (7.3) for $n = 3$:

[3] For $d = 3$, the result is trivial; the main interest lies in the cases with $d \geq 4$.

$$f(x_1, x_2, x_3) = \sum_{i=1}^{3} \Big\{ \phi_i[u_i(x_1, x_2), x_3] + \psi_i[v_i(x_1, x_2), x_3] + \chi_i[w_i(x_1, x_2), x_3]. \Big\}, \tag{7.4}$$

which means that any continuous function of three variables might be broken down to a sum of 9 functions ϕ_i, ψ_i, χ_i, $i = 1, 2, 3$, each of which being a single superposition of continuous functions of 2 variables. Furthermore, all properties of f are perfectly determined by those of the functions on the right-hand side of equations (7.3) and (7.4) [Arnold, 2009b, 1957].

Therefore, following the same wording as Pólya and Szegö, one may state that, *essentially, there is neither continuous nor discontinuous functions of three variables.* In addition, equation (7.4) reveals the incorrectness of Hilbert's conjecture in the 13th problem as well.

However, equation (7.4) was not in its "sharpest" form yet.[4] Apparently, Kolmogorov was not convinced that it was the end of the story. *What if there are no continuous functions of two variables either?* Or, what if any arbitrary continuous, bivariate function might be reduced to a superposition of continuous *univariate* functions?

7.1.3 Continuous Bivariate Functions Do Not Exist Either

There is a subtlety here which is specific to converting multivariate functions into univariate ones. The problem is that most of the basic arithmetic operations are, in general, functions of two variables. An example is the simple *addition*: $g(x, y) = x + y$.[5] So, we have to keep such nuances in mind and allow including simple arithmetic operations to our repository of univariate functions. However, later we will see that one of the beauties the next Kolmogorov's finding possesses is the fact that this violation of adding non-univariate functions to our repository can be kept, indeed, as minimum as possible.

First, we would like to, quickly, mention another result, also belonging to Kolmogorov, before moving onto the sharpest form of the representation theorem. In a sense, this result, also proved in 1956, is weaker than the final one. It states that any continuous, real-valued function, defined on I^d, $d \geq 2$, can be approximated by any desirable precision (in the sense of the *Chebyshev metric*) by means of superpositions of univariate polynomials and addition. Particularly, any continuous, bivariate function $f(x, y)$ can be approximated by a

[4] A. G. Vitushkin in Vitushkin [2004] mentions an interesting point about A. N. Kolmogorov: "It was in the character of Kolmogorov to carry any work to completion. Shortly thereafter, following the rule of improving every result to its sharpest form, ..."

[5] Here, we are interested in the general case, i.e., where x and y are unrelated to each other. Of course, if there is a relationship, say $y = \varphi(x)$, then $x+y$ happens to be a univariate function: $g(x, y) = x+y = x+\varphi(x)$.

superposition of the form

$$\sum_{i=1}^{2} P_i(y)\, Q[R_i(y) + x],$$

where P_i, Q, and R_is are suitable polynomials [Kolmogorov, 1961, Theorem 4].

From the mathematical analysis point of view, now, such a result is only of historical interest as, firstly, it is weaker than the final representation theorem, and secondly, it is only an "approximate" reply to the question of existence of continuous functions of two variables.

However, from the learning theory viewpoint, this is, conceptually, closer to what should be implemented in the context of the mapping neural networks, for example. This is a very important point and we will return to it later, in this chapter.

In fact, it did not take long for Kolmogorov to solve the problem completely and put an end to a series of works and studies, exclusively focused on continuous functions, invoked by Hilbert's thirteenth problem. In 1957, he proved the strongest possible form of *complexity reduction* for continuous functions in a nowadays-well-known result, commonly referred to as the *Kolmogorov-Arnold Representation/Superposition Theorem*:

3.1.2 Kolmogorov-Arnold Representation Theorem. *For any integer $d \geq 2$, there exists a family of certain continuous functions $g_{i,j} : I = [0, 1] \to \mathbb{R}$ such that any continuous function $f : I^d \to \mathbb{R}$ is representable in the form*

$$f(x_1, x_2, \ldots, x_d) = \sum_{i=1}^{2d+1} h_i\Big[\sum_{j=1}^{d} g_{i,j}(x_j)\Big], \tag{7.5}$$

with h_i's also being real-valued and continuous [Kolmogorov, 1991].

Now, consider the special case of $d = 3$, and let $\xi_i(u, v) = g_{i,1}(u) + g_{i,2}(v)$ and $F_i(u, v) = h_i\left[u + g_{i,3}(v)\right]$. Then, equation (7.5) implies that

$$f(x_1, x_2, x_3) = \sum_{i=1}^{7} F_i\left[\xi_i(x_1, x_2), x_3\right], \tag{7.6}$$

which is a slight improvement over equation (7.4), in terms of the total number of the terms used in the conversion. While in equation (7.4) there are 9 functions $\phi_i, \psi_i, \chi_i, i = 1, 2, 3$, each of which being a single superpositions of bivariate functions, in equation (7.6) there are only 7 such superpositions, i.e., $F_i, i = 1, \ldots, 7$.

Moreover, functions $g_{i,j}$'s in equation (7.5) are certain functions, whose properties are usually

well known. Note that $g_{i,j}$'s are independent from f and only depend on d. As a result, all properties of f might be thoroughly determined by those of the seven univariate functions h_i. (For more details see Kolmogorov [1991] and Arnold [2009b].)

Kolmogorov's proof of theorem 7.1.3 is rather elementary but extremely elegant at the same time. We shall skip the proof here since it does not serve the objective of the present work.

The following point is worth noting about Hilbert's thirteenth problem and the solution provided by the Kolmogorov-Arnold representation. This representation and the series of preceding works, specifically, dealt with *continuous* functions. As it was mentioned earlier, Hilbert's problem can be viewed from different perspectives. We shall briefly discuss some of these perspectives, in the current section, but in the meanwhile, we would like the reader to remember that, despite the incredibly powerful, inspiring, and elegant results that Kolmogorov and Arnold obtained, their relation to Hilbert's thirteenth problem should not be overestimated. One may consider the thirteenth problem in settings other than continuous functions, which might be important, also, from the practical point of view. What if stronger restrictions, rather than continuity, are imposed on the family of the target functions or those being used in the representation of the target functions? For example, providing a computation-friendly solution to Hilbert's 13th problem is yet to be studied.

Many of these aspects deserve investigation, especially, with respect to their applicability in learning theory. Perhaps, it would be no exaggeration to say that Hilbert's problem, in its entirety, is still as attractive problem as in the beginning of the twentieth century. Next, we briefly discuss some of the alternative settings, in which the thirteenth problem has been considered before.

To the best of our knowledge, until recently, all of the results considering more restrictive constraints (some of which being motivated by practice) on the family of functions to be employed in the decomposition of a multivariate function have approved Hilbert's conjecture. As mentioned before, Hilbert himself mentioned in the statement of the thirteenth problem that he had been aware of the *impossibility* of representing all *analytic* functions of 3 variables by superpositions of analytic bivariate functions. More general results concerning analytic functions were obtained by Ostrowski [1920], [6] who proved that a *bivariate analytic function of the particular form*

$$f(x,y) = \sum_{n=1}^{\infty} \frac{x^n}{n^y}$$

cannot be broken down to a finite superposition of infinitely differentiable univariate functions

[6] Unfortunately, an English translation of the article was not available to the author and Ostrowski's result mentioned here were extracted from Vitushkin and Khenkin [1967].

and algebraic functions of any number of variables.

Another related results belonging to Vitushkin [1954] illustrates that there exists a function of d variables, which is p-time differentiable, that *cannot* be represented as a finite superposition of d'-variate, p'-time differentiable functions provided that $\frac{d}{p} > \frac{d'}{p'}$. There are a couple of interesting points about this result: First, the quotient of the *number of variables over the degree of differentiability*, i.e., $\frac{d}{p}$, might be regarded as a *measure of complexity* for a family of functions. We shall elaborate on this shortly. Second, this result was obtained by applying the *theory of multidimensional variations*, developed by Vitushkin, however, later Kolmogorov arrived at the same result by *estimating the number of elements in covering ε-nets of the functional space of continuous and differentiable functions.* Let F_p^d denote the set of all continuous functions defined on the unit d-dimensional interval I^d, whose partial derivatives, up to order p inclusive, are all continuous and bounded by a constant c. Also, suppose that $N_{\varepsilon}(F_p^d)$ represents the minimum number of the elements in an ε-net covering the whole F_p^d. It turns out that

$$\lim_{\varepsilon \to 0} \frac{\log\{\log[N_{\varepsilon}(F_p^d)]\}}{\log(\frac{1}{\varepsilon})} = \frac{d}{p},$$

which, in fact, suggests that $\frac{d}{p}$ is equivalent to *ε-entropy* that is also a complexity measure of a functional space. We shall skip the details here but an interested reader may find an easy-to-follow explanation of the related issues in Tikhomirov [1963].

Now, as was promised earlier in this section, to grasp the intuition, at least roughly, behind the role of the ratio $\frac{d}{p}$ as a complexity measure for F_p^d, imagine, for example, $d = d'$ and then it should not be difficult to see why $\frac{d}{p} > \frac{d'}{p'}$ conveys, in some sense, that F_p^d is "massier" or "denser" than $F_{p'}^{d'}$. Roughly, when the number of arguments is fixed, to achieve a smoother function, one needs to impose more constraints on the function (in this case, smoothness constraints), and hence, fewer number of functions will be able to satisfy these additional conditions. [Kolmogorov, 1955, Vitushkin and Khenkin, 1967].

The next result, also belonging to Vitushkin, states that for arbitrary continuous functions $p_i(x_1, x_2, \ldots, x_d)$'s and continuously differentiable functions $q_i(x_1, x_2, \ldots, x_d)$, $d \geq 2$, $i = 1, 2, \ldots, N$, the set of superpositions of the form

$$\sum_{i=1}^{N} p_i(x_1, x_2, \ldots, x_d)\, f_i\left[q_i(x_1, x_2, \ldots, x_d)\right],$$

where f_i's are arbitrary univariate continuous functions, is nowhere dense in the space of all continuous functions of d variables [Vitushkin, 1964, Khenkin, 1964].

The last related result we would like to introduce here, dates back to 1951, a few years

before Kolmogorov and Arnold proved the famous superposition theorem. Alexey Alexeevich Milyutin, also a soviet mathematician, proved that *the linear space of continuous, univariate functions is isomorphic to the linear space of continuous functions of d variables, for an arbitrary positive integer d*. Furthermore, he succeeded to show that there does not exist such an isomorphism if, additionally, one assumes that the *smooth* functions of one space are restricted to be mapped only to those of the other space, i.e,. the smoothness of the functions is preserved under the purportedly isomorphism. Unluckily, the paper was not published because A. Pełczyński believed that the result was incorrect. Subsequently, this fascinating result was not recognized until fifteen years later when G. M. Khenkin arrived at the same result and learnt (possibly from Pełczyński) about Milyutin's work. His curiosity forced him to find the manuscript of Milyutin's work, and after reading it carefully he realized that the work was absolutely genuine and there were no mistakes! Khenkin himself then arranged the publication of Milyutin's paper in 1966 [Milyutin, 1966].

7.1.4 Generalizations of Kolmogorov's Superposition

Although Kolmogorov tried to bring the representation theorem to its sharpest form (which was accomplished to a reasonable extent), there have been several successful improvements of the Kolmogorov-Arnold representation theorem, carried out by other mathematicians.

These modifications can be roughly classified in two subgroups: First, those results that, technically, *simplified* the representation itself; that is, the right-hand side of equation (7.5). Notable examples are the results achieved by George Lorentz and David Sprecher in sixties and seventies, which considerably reduced the number of functions used in the superposition [Lorentz, 1962, Sprecher, 1965, 1966, 1972].

The second subgroup includes different *generalizations* of the theorem, usually extension to more general spaces. For example, Tikhomirov [1963] generalized the representation theorem to the product metric spaces. A similar result was later obtained by Ostrand [1965].

In the following subsection, we will discuss some of these adjusted versions of the superposition theorem with more details, because of their importance in building the theoretical connection between the Kolmogorov-Arnold representation theorem and the theory of perceptrons. But before moving further, we would like to conclude the current section by recalling that the topics discussed in this section, including their related questions, have been attracting an immense amount of attention over the past 150 years. Numerous mathematicians, as well as other researchers, have contributed to the solution of these questions or simply to extending our understanding of them, as a consequence of which, a lot of distinct perspectives have been involved and a vast variety of results have been generated. Therefore, the

author would like to acknowledge that the list of works, provided in this section, is far from being comprehensive for the following reasons: The main objective of the current section is merely to shed some light on the relation of a specific type of learning method, sometimes referred to as the *mapping neural networks*, with the connected results mostly belonging to mathematical analysis and approximation theory. We believe that the present condensed introduction suffices to connect the dots and achieve this end.

Another concern relates to the applicability of the Kolmogorov-Arnold representation theorem, particularly, in a neural network. This is a general concern Unfortunately, functions $g_{i,j}$'s can hardly be considered as "well-behaved" enough for practical purposes. Despite being continuous, they essentially suffer from a severe lack of smoothness. For instance, one such function, which is actually used in Kolmogorov's construction as one of $g_{i,j}$'s, is the Weierstrass function. Including such pathological functions makes this, theoretically, brilliant result almost useless in practice.

As mentioned before, there have appeared several modified versions of the Kolmogorov-Arnold Superposition theorem since its original introduction by Kolmogorov in 1957. Particularly, the first type of the modifications, i.e., those meant to simplify the structural form of the representation of the function f (see subsection 7.1.4), are essential for convenience and ease of implementation of the superposition theorem via networks. More precisely, it would be hard to associate any practical benefits to the implementation of a network based on the original statement of the Kolmogorov-Arnold representation theorem, without carrying out certain adjustments. Note that, from the theoretical point of view, it is still possible to introduce a perceptron with two internal layers that can compute exactly any continuous function $f : I^d \to \mathbb{R}$ at an arbitrary point of its domain by merely using the Kolmogorov-Arnold decomposition, nonetheless, as we will see later, realizing such a network in practice is far from being straightforward. In other words, while, theoretically speaking, one can prove the existence of the Kolmogorov-Arnold representation network, it is not clear if there is any general recipe to implement such a network for solving real-world problems.

In the following paragraphs, we will briefly introduce the improved versions of the representation theorem:

First, Lorentz [1962] showed that the outer functions $h_i, i = 1, \ldots, 2d+1$ in equation (7.5) might be replaced with a single function. Sprecher [1965] simplified it further by establishing that any function $f : I^d \to \mathbb{R}$ could be represented as

$$f(x_1, x_2, \ldots, x_d) = \sum_{i=1}^{2d+1} h\Big[\sum_{j=1}^{d} \lambda^{ij} g(x_j + i\epsilon)\Big], \tag{7.7}$$

where $h : \mathbb{R} \to \mathbb{R}$ is a continuous function, λ and ϵ are constants, $g : \mathbb{R} \to \mathbb{R}$ is an increasing, Hölder-continuous function. None of λ, ϵ, and g depends on f (see also Sprecher [1966]).

Two years later, Fridman [1967] showed that the inner functions in equation (7.5) could be chosen to be Lipschitz-continuous, however, the number of the inner and outer functions were identical to that of the original theorem by Kolmogorov. Eventually, in 1972, Sprecher himself succeeded to improve his initial representation by providing a new formula with *smoother* inner functions as follows: [Sprecher, 1972]

Sprecher's Representation. *For any integer $d \geq 2$, there exists a constant $\lambda \neq 0$, and a function $g \in Lip(1)$ such that any continuous function $f : I^d \to \mathbb{R}$ might be represented in the form*

$$f(x_1, x_2, \ldots, x_d) = \sum_{i=1}^{2d+1} h_i\Big[\sum_{j=1}^{d} \lambda^{j-1} g(x_j + i\epsilon)\Big], \tag{7.8}$$

where h_i's are continuous real-valued functions and ϵ is any non-zero constant.

Note that the functions h_i can be replaced with a single continuous function h by adding a suitable constant to each $\xi_j = \sum_{j=1}^{d} \lambda^{j-1} g(x_j + i\epsilon)$ [Sprecher, 1972, 1993].

Equations (7.7) and (7.8) provide relatively convenient settings for the approximation of the function f through multilayer perceptrons. To see that, in the next subsection, we first give a general account of the creation of the neural networks and explain the procedure based on which a neural network functions. Then, in the next step, the adaptation of the neural networks for estimating functional dependencies according to representation theorems will be discussed.

7.2 Function Estimation by Neural Networks

The ultimate goal of this section is to trace the events leading up to the theories that connect the Kolmogorov-Arnold superposition theorem and its later-adjusted versions with *multilayer perceptrons* or *neural networks*. To this end, we have to consider various theoretical results belonging to different branches of mathematics and learning theory. The creation and development of the *theory* of neural networks has been significantly benefited from achievements in several domains of mathematics. A considerable amount of these theories were developed during 1960s and 1970s. To understand the interconnection amongst these results and, in particular, with *multilayer perceptrons* (also, called *neural networks*), we will discuss them in

parallel.

7.2.1 The First Learning Machine Was A Neural Model

In late 1950s and 1960s, almost at the same time as the superposition theorem was being built and burnished, Frank Rosenblatt, an American psychologist, was busy with the creation of the *first learning machine*, i.e., Rosenblatt's *perceptron*.[7] As a theoretical model, the idea of perceptron had been around for years and was utilized for describing the process of learning in living organisms. However, the novelty of Rosenblatt's perceptron was in its implementation as a computer program, which makes it the first artificial neural network came to existence.[8] The perceptron was designed based on the McCulloch-Pitts model of a single neuron, which takes a vector of input values and produces a binary output according to a simple relation

$$y = \text{sgn}[\langle \mathbf{w}, \mathbf{x} \rangle - b],$$

where $\mathbf{w} = (w_1, \ldots, w_d), \mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ denote a vector of weights and inputs, respectively, $b \in \mathbb{R}$ is a threshold value, and $y \in \{-1, +1\}$ is the output (Figure 7.2). This is perhaps the simplest model of pattern recognition learning machine that employs a "caricature" model of a single unit to solve a biclassification problem.

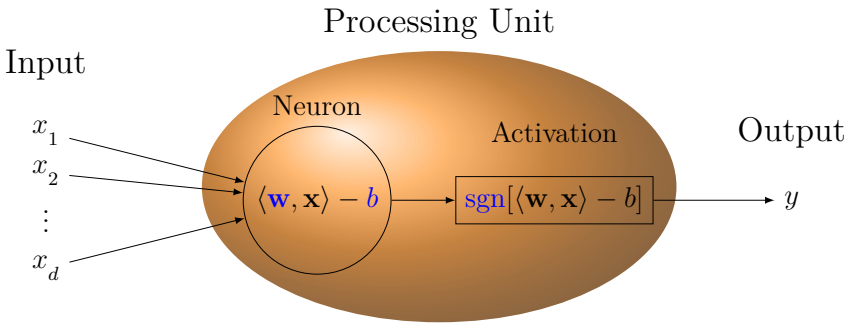

Figure 7.2 **Rosenblatt's Neural Unit.** This figure shows a processing unit of Rosenblatt's perceptron based on the McCulloch-Pitts neuron model.

Let $\mathcal{X} \subset \mathbb{R}^d$ be the set of all possible input vectors and define

$$h(\mathbf{x}, \mathbf{w}) := \langle \mathbf{x}, \mathbf{w} \rangle - b.$$

[7] In fact, the systematic study of the problem of learning was born only after the advent of the first learning machine by Rosenblatt.

[8] The term artificial neural network, however, appeared some years later for referring to a more complex type of learning machine also inspired by the brain's neural procedure of learning.

Geometrically speaking, any point of $\mathcal{X}$, not belonging to the hyperplane $h(\mathbf{x}, \mathbf{w}) = 0$, falls into either of the following two subsets of the input space $\mathcal{X}$:

$$\mathcal{X}_{+} := \{\mathbf{x} \in \mathcal{X} \mid h(\mathbf{x}, \mathbf{w}) > 0\}$$
$$\mathcal{X}_{-} := \{\mathbf{x} \in \mathcal{X} \mid h(\mathbf{x}, \mathbf{w}) < 0\},$$

for a fixed vector $\mathbf{w}$. Obviously, $\mathcal{X}_{+}$ and $\mathcal{X}_{-}$ correspond to the previously mentioned outputs $y = +1$ and $y = -1$, respectively. This is a simple *pattern recognition* problem and, hence, a single unit is able to classify a set of vectors with binary labels.

Rosenblatt's perceptron consisted of multiple units divided into several layers such that the outputs of each layer were passed to another layer of units as inputs. The output of each unit might have been transferred to multiple units in the next layer. The last layer, i.e., the output layer, however had a single unit which generated the final label. See Figure 7.3 for an example of a perceptron.

Learning in the context of the perceptron consists of finding an optimal set of values for all weights $\mathbf{w}$ and thresholds b for each single neuron. Back in sixties it was not known how to determine all those values, simultaneously, as *backpropagation* was introduced more than two decades later. Rosenblatt's solution to this problem was, first, fixing all the weights and thresholds of the network except for those of the last neuron, which would be learnt from data in an iterative manner. We will skip the details of the training procedure in Rosenblatt's perceptron for a reason we will state shortly, however the reader is referred to Vapnik [1995] for a brief discussion, and to Rosenblatt [1962] for a full-length description. The reason we have mentioned the perceptron model here is merely to emphasize that "learning from examples" (or shortly, "learning"), as is understood in today's modern theory of learning was happening in Rosenblatt's perceptron although partly.

While the perceptron might be regarded as the first actual learning machine, the mathematical analysis of learning was born, shortly after, just by the work of Novikoff in 1962. He proved an upper bound for the number of mistakes the perceptron makes until it learns the separating hyperplane. In other words, he showed that if the data is perfectly separable by margin $\rho > 0$ and all input vectors are bounded, i.e., for every vector $\mathbf{x}$ in the sample data it holds that $\|\mathbf{x}\| \leq R$, for some positive real R, then the maximum number of coefficient corrections needed to be undertaken by the model in order to discover the optimal values of the weights and thresholds is less than or equal to $\frac{R^2}{\rho^2}$ [Novikoff, 1962, 1963].

Despite the early promising achievements in the context of neural networks, several serious limitations in terms of applicability of the perceptron remained unsolved. As mentioned

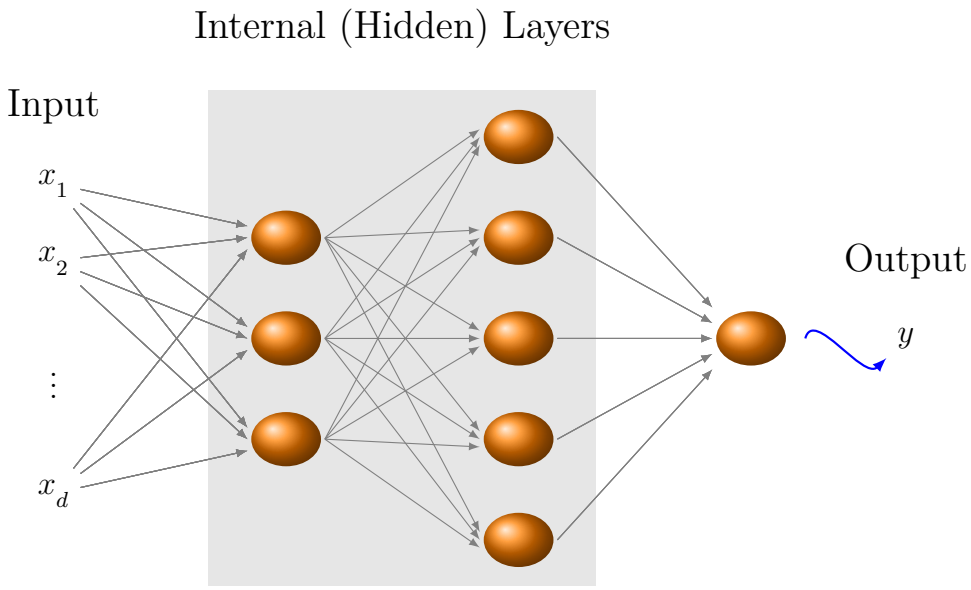

Figure 7.3 **Rosenblatt's Perceptron with Two Internal Layers.** An example of Rosenblatt's perceptron is illustrated. This perceptron contains two internal (hidden) layers. The 1st one consists of three processing units, while the 2nd one has five units. This perceptron processes a d-dimensional input vector $\mathbf{x}$ and returns a real-valued output y.

earlier, particularly, a general procedure for determining the coefficients of the network was yet to be found. The years following the emergence of the perceptron witnessed some advances in the context of the neural networks but the next influential moment in the history of the neural networks was probably the *popularization* of the gradient-based fitting methods, called *backpropagation*, by Rumelhart et al. [1986b,a]. The invention of the backpropagation, however, dates back to 1960s in control theory [Kelley, 1960, Bryson, 1962].[9] The first computer implementation of it was done by Linnainmaa [1970].

Note that with the activation function originally used in perceptron, i.e., the *sign* function it was not possible to use gradient-based methods because of non-differentiability. This problem was overcome in backpropagation by approximating the sign function by means of a sigmoidal function such as hyperbolic tangent. Introduction of the sigmoidal functions as the activation function, in fact, paved the way for using any gradient-based optimization technique such as *gradient descent*.

Following the popularization of the backpropagation in the context of neural networks, scientists' interest in application of networks in different fields began to rise for the second time. In fact, many researchers regard the second half of 1980s as the actual birth date of the

[9] Backpropagation was rediscovered, independently, by several people during 1960s and 1970s [Dreyfus, 1962, Linnainmaa, 1976, Parker, 1985, Werbos, 1994, Vapnik, 1995, Goodfellow et al., 2016].

profound theoretical development of the neural networks. It was also at that time that the American computer scientist, Hecht-Nielsen, pointed out for the first time the resemblance of the machinery used for function approximation in the Kolmogorov-Arnold representation theorem and the structure of a multilayer perceptron in Hecht-Nielsen [1987]. He interpreted the improved version of the representation by Sprecher as a three-layer perceptron whose activation functions were the inner and outer functions of the superposition. In an interesting statement, in Hecht-Nielsen [1987], where he proved the existence of Kolmogorov's mapping neural network, he claimed that "in mathematical terms, no one has found a significant use for it [Kolmogorov's theorem]. The point of this paper is that this is *not* the case in neurocomputing!" Nevertheless, Hecht-Nielsen himself expressed his dissatisfaction with the theorem being nonconstructive, that is, while it guarantees the existence of such a network, it does not provide a way to determine the outer functions: "The direct usefulness of this result is doubtful, at least in the near term, ..."

Similar objections also came from other researchers. Girosi and Poggio [1989] regarded the Kolmogorov-Arnold superposition theorem as irrelevant due to the inner and outer functions being highly non-smooth, which makes them very difficult to be approximated, as well as the outer function being dependent on the target function f. Later Kurková [1991] showed that this problem can be overcome by approximating both the inner and outer functions by means of specific combinations of sigmoidal functions.

Definition 14 (Sigmoidal Function)**.** *A continuous function $\sigma : \mathbb{R} \to I$ is said to be sigmoidal if the following two conditions hold:*

1. $\lim_{x \to -\infty} \sigma(x) = 0$, *and*
2. $\lim_{x \to \infty} \sigma(x) = 1$.

Kurková [1991] proved that the inner and outer functions can be approximated with an arbitrary precision utilizing the following linear combinations of a sigmoidal function:

$$\sum_{i=1}^{k} a_i \sigma(b_i x + c_i),$$

where $a_i, b_i, c_i \in \mathbb{R}$, $i = 1, \ldots, k$. This implies that any ordinary perceptron with the standard sigmoidal functions might be used. Moreover, Kurková provided the exact number of units needed in each hidden layer to obtain an approximation with the desired accuracy [Kurková, 1991, 1992].

A key technical point in the evolution of the mapping neural networks theory is turning attention from *representation* to *approximation*. That is, all versions of the Kolmogorov-Arnold representation theorem give an *exact* representation of the function f at any point $\mathbf{x}$ in its domain. In contrast, Kurková talks about approximating the inner and outer functions, which in turn leads to an approximation of $f(\mathbf{x})$. This rotation is more than just a technical adjustment: Considering the approximation question opens a new horizon in studying the mapping neural networks whose central aim is to investigate the approximation abilities of the neural network. Conceptually, it is a crucial turn from the learning theory perspective as it is more aligned with the objective of the inverse problem of *learning from examples* opposed to exact representation of a function which seems to be of more interest for mathematical analysis.

Studying the approximation capability of multilayer perceptrons, nevertheless, had begun before the aforementioned works of Kurková in $1991, 1992$. There exist several results, commonly referred to as the *universal approximation theorems*, that address the approximation abilities of multilayer feedforward neural networks with a variety of different conditions. Some are stronger than the others in the sense that they imply the other ones, while there are other ones that are not comparable and hold for different contexts. Below, we will give only the ones that we believe are more relevant to our work:

Some of the earlier results are due to Cybenko and Hornik as follows. Cybenko [1988, 1989] settled the question of the approximation ability of the neural networks with two and one internal layer, respectively. In particular, the latter demonstrated that any continuous function $f : I^d \to \mathbb{R}$ can be arbitrarily well approximated, with respect to the uniform norm, by a neural network with only one internal layer and a sigmoidal function as its activation function. As a consequence, it was established that arbitrary decision regions, i.e., any collection of compact, disjoint subsets of $\mathbb{R}^d$, might be discriminated to a desired degree of precision by means of continuous feedforward neural networks with solely one single internal layer and any continuous sigmoidal activation function.

Surprisingly, at the same time but independently, a similar result to that of Cybenko [1989] was obtained by Hornik et al. [1988, 1989]. However, they used very different approaches in their proofs: while Cybenko made use of the Hahn-Banach theorem, Hornik proved it applying the Stone-Weierstrass approximation theorem. Later, Hornik [1991] proved two extended versions of his former result by replacing the sigmoidal activation function with any *bounded* and *nonconstant* one. Let $\mathcal{X} \subseteq \mathbb{R}^d$ and μ be a finite measure defined on $\mathcal{X}$.

Also, let p be a positive real constant, and $L^p(\mathcal{X})$ be as follows:

$$L^p(\mathcal{X}) = \left\{ f : \mathcal{X} \to \mathbb{R} \;\middle|\; \int_{\mathcal{X}} |f|^p \, \mathrm{d}\mu < \infty \right\}.$$

Hornik [1991] concluded the next two statements:

1. Any standard multilayer feedforward network with a single internal layer and arbitrary *bounded* and *nonconstant* activation function is a universal approximator for an arbitrary function $f \in L^p(\mathcal{X})$, with respect to the corresponding L^p-norm, provided that a sufficient number of units are available.

2. Any standard multilayer feedforward network with a single internal layer and arbitrary *continuous*, *bounded* and *nonconstant* activation function is a *uniformly* universal approximator for an arbitrary function $f \in L^p(\mathcal{X})$, where $\mathcal{X}$ is a compact set, provided that a sufficient number of units are available.

Additionally, Hornik [1991] studied general conditions necessary for ensuring that networks with sufficiently smooth activation functions are capable of arbitrarily accurate approximation to a function and its derivatives.

Eventually, Leshno et al. [1993] succeeded to prove a vigorous extension of the result obtained by Hornik [1991], which elegantly established the *necessary and sufficient* conditions for the universal approximation property of multilayer perceptrons. The following definitions are needed for the precise formulation of the theorem.

Note that in the following paragraphs, μ is a Lebesgue measure defined on the set $\mathcal{X} \subseteq \mathbb{R}^d$. In addition, for any function f, the set of all its discontinuity points will be denoted by $Disc(f)$.

Definition 15 (Essentially Bounded Function)**.** *A real-valued function f defined almost everywhere, with respect to μ, on $\mathcal{X}$ is said to be essentially bounded on $\mathcal{X}$, if $|f(\mathbf{x})|$ is bounded almost everywhere on $\mathcal{X}$. The set of all essentially bounded functions on $\mathcal{X}$ is denoted by $L^\infty(\mathcal{X})$ and is equipped with the norm defined as*

$$\|f\|_{L^\infty(\mathcal{X})} := \inf \left\{ \lambda \mid \mu \{ \mathbf{x} : |f(\mathbf{x})| \geq \lambda \} = 0 \right\}.$$

Definition 16 (Locally Essentially Bounded Function)**.** *A real-valued function f defined almost everywhere, with respect to μ, on an open set $\mathcal{X}$ is said to be locally essentially bounded on $\mathcal{X}$, if for any point $\mathbf{x} \in \mathcal{X}$ there exists a compact set K such that $\mathbf{x} \in K$ and $f \in L^\infty(K)$. The set of all such functions is denoted by $L^\infty_{loc}(\mathcal{X})$.*

Necessity and Sufficiency for Universal Approximation [Leshno et al., 1993]. *Let Σ be the set of real-valued functions σ for which hold the following:*

(i) $\Sigma \subseteq L^{\infty}_{loc}(\mathbb{R})$, *and*

(ii) $\mu(\overline{Disc(\varphi)}) = 0$, *for any* $\varphi \in \Sigma$.

Let $\sigma \in \Sigma$ *and*

$$\Sigma_{*} = span\left\{\sigma\left(\langle \mathbf{w}, \mathbf{x}\rangle + \theta\right) : \mathbf{w} \in \mathbb{R}^{d}, \theta \in \mathbb{R}\right\}.$$

Then, Σ_{*} *is dense in* $C(\mathbb{R}^{d})$, *if and only if* σ *is non-polynomial.*

7.2.2 Numerical Implementation

Clearly, all the aforementioned results on the representation, and universal approximation theorems are purely from the theoretical point of view, whose practical usefulness, undoubtedly, depends on the implementation algorithms used. However, since in the present discussion, detailed numerical and programming aspects of the problem is not our primary concern we restrict the scope of the discussion in what follows to only mentioning some major attempts, almost all of which belong to the domain of shallow neural networks.

In 1990s numerical algorithms for implementing mapping neural networks based on the adjusted versions of the Kolmogorov-Arnold superposition started to appear. Sprecher [1996] provides a numerical implementation of the inner function based on the network proposed by Hecht-Nielsen [1987].[10] The numerical implementation of the outer function was accomplished in Sprecher [1997].

Igelnik and Parikh [2003] made use of cubic splines for approximation of the both inner and outer functions of Kolmogorov's superposition. The authors claimed that using splines could lead to the network architecture's increased capability of adapting itself to data. The called the network *Kolmogorov's Spline Network*.

Braun and Griebel [2009] provided a constructive proof for the Kolmogorov-Arnold superposition theorem, which used Hölder-continuous inner functions. They, also, proposed an algorithm for implementing it. (Also, see Braun [2009].)

In recent years, there has been a lot of attention devoted to the *deep neural networks*, especially, after they emerged triumphant in numerous applications. As a result of this new

[10] Recall that Hecht-Nielsen existence theorem itself was based on earlier Sprecher's improved version of the superposition theorem.

trend, some researchers began to exhibit interest in studying approximation capability and computational complexity of deep neural networks alongside other aspects of deep networks theory. One attractive question, for instance, regards the relation between the dimension of the input space, size of the network, and the approximation accuracy of a deep neural network. This problem has close connections with the notorious phenomenon called *curse of dimensionality*. While studying the theory of the deep neural networks is well beyond the scope of this thesis, below is provided a very short list of some related works, which we believe can serve as a good point to start for an interested reader: Ait Gougam et al. [2008], Eldan and Shamir [2016], Poggio et al. [2017], Yarotsky [2018], Petersen and Voigtlaender [2018], Liu and Liang [2019], Ohn and Kim [2019], Chen et al. [2019], Montanelli and Yang [2020].

CHAPTER 8 CONCLUSION AND RECOMMENDATIONS

8.1 Summary

This work has been dedicated to studying statistical learning from length-biased, right-censored, with covariates (LBRC-C) data. To this end, we had to, first, review the foundations of statistical learning theory, which was accomplished in Chapter 3. In particular, it was established that the inferential infrastructure of statistical learning theory is built based on an inductive principle of learning, called risk minimization. Two aspects of risk minimization problem, i.e., the empirical risk minimization (ERM) and structural risk minimization (SRM), were discussed. Finally, the necessary and sufficient conditions for reliable learning were given.

In Chapter 4, foundations of weakly supervised learning from LBRC-C data were established. Specifically, the following three major problems, in connection with biased and censored data, were settled: First, the problem of learning the distribution function; second, the problem of risk minimization, which indeed is the core of statistical learning theory, and finally, estimating the non-explicit regression function in presence of length bias and right censoring. All the aforementioned problems were solved to a satisfactory degree, although there are some aspects that need further investigation in future.

In Chapter 5, we studied the problem of variable selection in the context of regression analysis of LBRC-C data. Particularly, the implications of applying a theoretically-correct likelihood function, i.e., the joint likelihood function, for variable selection were investigated. It was discussed that when data are collected through a length-biased sampling scheme, then covariates suffer from an additional level of induced biased, which is imposed by the length bias itself. It was hypothesized that due to ignoring this bias by the conditional approach, one may end up with more incorrect models if likelihood-based selection criteria were to be used. Although a thorough mathematical proof, either in favour of or against this hypothesis, cannot be provided at the moment, a simulation study, provided in Chapter 6, supported it.

In the end, we surveyed the so-called mapping neural networks and their mathematical roots, in order to figure out their capability of solving the main problem of statistical learning or the problem of function estimation. It was observed that while the mapping neural networks might be very strong estimators of a reasonably broad class of functions, they are not able to solve the learning problem completely. In other words, like the classical parametric paradigm of statistical inference, neural networks have no mechanism to control the model's complexity

(capacity) in accordance with the data in hand. As discussed in Chapter 3, capacity control is a key factor in providing the reliability of a learning machine. As a result, the networks' reliability for learning cannot be theoretically guaranteed unless a considerable amount of prior information is available, based on which the network's complexity might be regulated. In practice, adjustments for complexity must be mainly done through "smart" heuristic methods.

8.2 Some Challenges and Future Research

This section addresses a few questions, a couple of which, in the opinion of the authors, compose some of today's most challenging problems of data science and machine learning. The majority of the discussions in this chapter are rather raw ideas and, therefore, require further investigation.

8.2.1 Classification Under Length Bias and Censoring

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 the electronic health records (EHR), including length bias and censoring, most well-known learning techniques cannot be applied naively. Several ad hoc approaches have been tried previously to adapt some machine learning techniques to the electronic health records (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 Vock et al. [2016]). 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., Ishwaran et al. [2008], Shivaswamy et al. [2007], Khan and Zubek [2008], Goldberg and Kosorok [2017], Luck et al. [2018], among others.

What is mostly missing in the literature is difficulties induced by left truncation. Hence, the classification problem we are interested in is as follows: Suppose that we are provided with length-biased, right-censored sample data $\mathcal{D}$ of the form

$$\mathcal{D} = \{(\tilde{A}_i, \tilde{R}_i \wedge C_i, \delta_i, \overset{*}{\mathbf{X}}_i) : i = 1, \ldots, n\}.$$

Also, assume that $\alpha > 0$ is some real. Then, we would like to evaluate the stochastic indicator function $\mathbb{1}_{\{Y \geq \alpha | A, \mathbf{X}\}}(a, \mathbf{x})$ at any point $(a, \mathbf{x})$. In other words, the goal is to predict whether a certain terminating event will happen in a certain amount of time for any new subject, given their current lifetime and covariates.

Taking proper care of induced covariate bias might be a crucial factor in methods, whose performance essentially relies on the characteristics of the input space. Examples of such methods are all kinds of tree-based methods. The classification problem described above is a special type of classification, i.e., a *binary classification* problem (also called *biclassification*) as the response is a binary variable. Nevertheless, the generalization to multiple classes should be straightforward. Once more, we would like to remind the reader that the novelty of the method is in treating the length bias and the covariate bias induced by the sampling scheme.

8.2.2 Causal Inference: Statistics vs. Machine Learning

The importance of *causation* lies in the fact that if one can rationalize phenomena in terms of *cause* and *effect*, then it becomes possible to change the outcome by modifying the cause. This is the core motivation of developing tools for *causal inference* in both statistics and machine learning.

Causal inference has always been a major component of statistical inference since its emergence. Before Fisher, the acceptable standard for inferring causal effects was conducting *controlled* experiments, where researchers tried to minimize the differences between the *treatment* and *control* arms as much as possible. However, in practice, it is very hard to achieve this goal in most of the cases, as a result of which the effects of the treatment may not be properly identified. Fisher was the first to come up with the idea of employing randomization in assigning units to either control or treatment groups. Theoretically, this maximizes the resemblance between the arms and cancels potential associations between the result of the treatment and the grouping [Fisher, 1935].

The points that we would like to highlight about the causal inference in the aforementioned statistical context are as follows: First, the *randomized controlled trials (RCTs)* are clearly designed for *experimental studies* and cannot be used in *observational studies*. Second, the problem is typically set in the language of testing hypothesis with the potential cause and effect being presumed before the experiment starts. These points are important to us since it turns out that they comprise, indeed, the primary distinction between the causal inference in statistics and machine learning.

Compared to classical statistics, the problem of causality became an issue more recently. Nevertheless, its vital importance has been acknowledged by several pioneering figures in the domain [Pearl and Mackenzie, 2018, Schölkopf, 2019, Bengio, 2019, Peters et al., 2017]. The reason why causality is of interest in machine learning is that it might pave the way to creation of the learning machines which can generalize the learnt skills from one task to

another, without the need of being trained from scratch. This is probably the next step in getting closer to *actual* artificial intelligence.

Interestingly, the problem of causality in the learning context is substantially different from what we described in statistics. Namely, suppose that a sample set containing n realized values of the random variables X and Y is given, i.e., $\{(x_i, y_i) : i = 1, 2, \ldots, n\}$. The causal problem is defined as to identify whether X causes Y or vice versa. In other words, the problem of causation in machine learning involves detecting the right direction of the causal relationship.

Before moving onto more detailed description of the problem, we would like to, swiftly, bring a couple of subtle points to the reader's attention. It is, also, worth mentioning that, although such concerns might be regarded as of minor importance from the microscopic view, they should not be ignored if the conceptual integrity of the theory is a concern.

The first issue is that, in contrast to the classical framework of causal inference in statistics, seemingly, the problem of causality has been defined and understood differently by different authors. In other words, it is difficult to give a consensual understanding of the causality problem that is consistent throughout the literature. The second problem relates to the conceptual legitimacy of the "causality" or "causation" being considered in some sources. (See, e.g., the discussion on the relevance of *time* to causality in Peters et al. [2017]). Without entering the details, we would like to clarify that, in the terminology adopted for the remainder, instead of "causality", the terms *explainability*, *precedence* or *dependency* will be used depending on the concept being addressed.

One of the approaches being used in order to identify the *precedence direction*, between X and Y, is by comparing the variation between the marginal distribution of one of the variables with the conditional distribution of the other variable given the first one. However, it is possible to show that covariation is not *generally* capable of detecting the direction of the dependency between two random variables. Consider the following regression:

Let $Y = a + bX + \varepsilon$, where $a, b \in \mathbb{R}$, $\varepsilon \sim N(0, \sigma_\varepsilon^2)$, and $X \sim N(0, 1)$. That is, $Y_{|X=x} \sim N(a + bx, \sigma_\varepsilon^2)$. Then,

$$Y \sim N(a, \sigma_\varepsilon^2 + b^2)$$

and

$$X_{|Y=y} \sim N\left(\frac{b(y-a)}{\sigma_\varepsilon^2 + b^2}, \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + b^2}\right).$$

Nevertheless, we have that

$$\frac{\mathrm{Var}(Y) - \mathrm{Var}(Y|X=x)}{\mathrm{Var}(Y)} = \frac{\mathrm{Var}(X) - \mathrm{Var}(X|Y=y)}{\mathrm{Var}(X)} = \frac{b^2}{\sigma_\varepsilon^2 + b^2},$$

which, clearly, means that variation cannot distinguish the precedence, i.e., $X \to Y$ from $Y \to X$. Note that X is ancillary for parameters of $Y|X$; X does not contain information about the parameters of $Y|X$. In contrast, Y clearly contains information about the parameters of $X|Y$. Therefore, precedence can be captured through information in X about $Y|X$ and information in Y about $X|Y$:

$$\frac{\partial}{\partial \theta_i} \log p_{\boldsymbol{\theta}}(x) = 0, \quad \text{for } i = 1, 2, 3,$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3) = (a, b, \sigma_\varepsilon^2)$. In our example, we have, in fact,

$$\frac{\partial}{\partial \theta_i} \log p_{\vec{\theta}}(x, y) = \frac{\partial}{\partial \theta_i} \log p_{\vec{\theta}}(y|x), \quad \text{for } i = 1, 2, 3. \tag{8.1}$$

Now, recall the length-biased sampling design considered in the previous chapter: The sample at disposal is not a representative sample of the population; in particular, the sampling distribution of the covariates is biased. Consequently, one cannot rely on the information contained in X about the conditional distribution of $Y|X$. In such a case equation (8.1) fails to hold, even if $X \to Y$.

In conclusion, the problem of identifying the direction of precedence or dependency in the aforementioned setting remains open and requires further investigation.

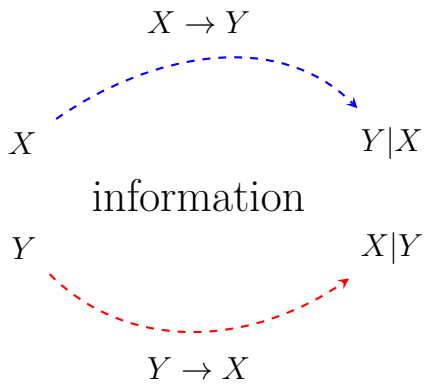

Figure 8.1 **Precedence Detection by Information Content.**

8.2.3 Intrinsic Dimension

Today, availability of huge data is constantly becoming less of a privilege, enjoying which is solely restricted to governments and few extremely reach organizations. Due to recent technology, particularly, the advent of advanced electronic hardware, storing data with thousands or millions of variables is not a big deal anymore.

Therefore, developing techniques for big data analysis is a growing necessity in the related fields. This need has been addressed in many different but related ways during the past several decades. For instance, the very well-studied techniques of variable selection, feature extraction methods of manifold learning, multidimensional scaling techniques, linear and nonlinear methods of dimensionality reduction, and etc are all responses to this necessity from different perspectives.

Despite having a lot of practical techniques for dealing with high-dimensional data in practice, there is still a considerable amount of questions left to be answered. The very first question is, does there exist a unifying theoretical framework that can formally define the intrinsic dimensionality of data? Another important question is, what are other possible factors that comprise the complexity of data? Or, how the intrinsic dimension of data is interrelated with the complexity of a learning machine or a statistical model? From the practical point of view, it would be phenomenally interesting and useful to design a learning machine that is able to learn the intrinsic dimension of data in practice. And finally, one may ask all the aforementioned questions in relation with biased data.

REFERENCES

V. N. Vapnik. *Statistical Learning Theory*. JOHN WILEY & SONS, INC., 1998. ISBN 978-0-471-03003-4. URL https://www.wiley.com/en-ca/Statistical+Learning+Theory-p-9780471030034.

Benjamin D. Horne, Heidi T. May, Joseph B. Muhlestein, Brianna S. Ronnow, Donald L. Lappé, Dale G. Renlund, Abdallah G. Kfoury, John F. Carlquist, Patrick W. Fisher, Robert R. Pearson, Tami L. Bair, and Jeffrey L. Anderson. Exceptional Mortality Prediction by Risk Scores from Common Laboratory Tests. *The American Journal of Medicine*, 122 (6):550–558, June 2009. ISSN 0002-9343, 1555-7162. doi: 10.1016/j.amjmed.2008.10.043. URL https://www.amjmed.com/article/S0002-9343(09)00103-X/abstract. Publisher: Elsevier.

Jenni A. M. Sidey-Gibbons and Chris J. Sidey-Gibbons. Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology*, 19(1):64, March 2019. ISSN 1471-2288. doi: 10.1186/s12874-019-0681-4. URL https://doi.org/10.1186/s12874-019-0681-4.

Michael E. Matheny, Danielle Whicher, and Sonoo Thadaney Israni. Artificial Intelligence in Health Care: A Report From the National Academy of Medicine. *JAMA*, 323(6):509–510, February 2020. ISSN 0098-7484. doi: 10.1001/jama.2019.21579. URL https://doi.org/10.1001/jama.2019.21579.

Peter F. Halpin and Henderikus J. Stam. Inductive Inference or Inductive Behavior: Fisher and Neyman: Pearson Approaches to Statistical Testing in Psychological Research (1940-1960). *The American Journal of Psychology*, 119(4):625, December 2006. ISSN 00029556. doi: 10.2307/20445367. URL https://www.jstor.org/stable/10.2307/20445367?origin=crossref.

Sir Ronald Aylmer Fisher. *The Design of Experiments*. Oliver and Boyd, London, 1935.

Marc J. LeClere. PREFACE Modeling Time to Event: Applications of Survival Analysis in Accounting, Economics and Finance. *Review of Accounting and Finance*, 4(4):5–12, January 2005. ISSN 1475-7702. doi: 10.1108/eb043434. URL https://doi.org/10.1108/eb043434. Publisher: Emerald Group Publishing Limited.

Steven Backman, Andrew Baker, Scott Beattie, Penelope Brasher, Gregory Bryson, Davy Cheng, Mark Crawford, Alain Deschamps, Francois Donati, Pierre Drolet, et al. 2011

canadian journal of anesthesia guide for authors. *Canadian Journal of Anesthesia/Journal canadien d'anesthésie*, 58(7):668–696, 2011.

Lucy Asher, Naomi D. Harvey, Martin Green, and Gary C. W. England. Application of Survival Analysis and Multistate Modeling to Understand Animal Behavior: Examples from Guide Dogs. *Frontiers in Veterinary Science*, 4, 2017. ISSN 2297-1769. doi: 10.3389/fvets.2017.00116. URL https://www.frontiersin.org/articles/10.3389/fvets.2017.00116/full. Publisher: Frontiers.

Shinichi Nakagawa and Robert P. Freckleton. Missing Inaction: The Dangers of Ignoring Missing Data. *Trends in Ecology & Evolution*, 23(11):592–596, November 2008. ISSN 0169-5347. doi: 10.1016/j.tree.2008.06.014. URL https://www.sciencedirect.com/science/article/pii/S0169534708002772.

Mei-Cheng Wang, Ron Brookmeyer, and Nicholas P. Jewell. Statistical Models for Prevalent Cohort Data. *Biometrics*, 49(1):1–11, 1993. ISSN 0006-341X. doi: 10.2307/2532597. URL http://www.jstor.org/stable/2532597. Publisher: [Wiley, International Biometric Society].

Masoud Asgharian and David B. Wolfson. Asymptotic Behavior of the Unconditional NPMLE of the Length-Biased Survivor Function from Right-Censored Prevalent Cohort Data. *Annals of Statistics*, 33(5):2109–2131, 2005. ISSN 0090-5364, 2168-8966. doi: 10.1214/009053605000000372. URL https://projecteuclid.org/euclid.aos/1132936558. Publisher: Institute of Mathematical Statistics.

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. doi: 10.1198/016214508000000382. URL https://doi.org/10.1198/016214508000000382.

Christina Wolfson, David B Wolfson, Masoud Asgharian, Cyr Emile M'Lan, Truls Østbye, Kenneth Rockwood, and DB ft Hogan. A reevaluation of the duration of survival after the onset of dementia. *New England Journal of Medicine*, 344(15):1111–1116, 2001.

Stephen W Lagakos, Leila M BARRAJ, and V de Gruttola. Nonparametric analysis of truncated survival data, with application to aids. *Biometrika*, 75(3):515–523, 1988.

Kwan-Moon Leung, Robert M. Elashoff, and Abdelmonem A. Afifi. Censoring Issues in Survival Analysis. *Annual Review of Public Health*, 18(1):83–104, 1997. doi: 10.1146/annurev.publhealth.18.1.83.

URL https://doi.org/10.1146/annurev.publhealth.18.1.83. _eprint: https://doi.org/10.1146/annurev.publhealth.18.1.83.

Enrique Barrajon and Laura Barrajon. Effect of Right Censoring Bias on Survival Analysis. *Journal of Clinical Oncology*, 37(15_suppl):e18188–e18188, May 2019. ISSN 0732-183X. doi: 10.1200/JCO.2019.37.15_suppl.e18188. URL https://ascopubs.org/doi/abs/10.1200/JCO.2019.37.15_suppl.e18188. Publisher: Wolters Kluwer.

S. W. Lagakos. General Right Censoring and Its Impact on the Analysis of Survival Data. *Biometrics*, 35(1):139–156, 1979. ISSN 0006-341X. doi: 10.2307/2529941. URL http://www.jstor.org/stable/2529941. Publisher: [Wiley, International Biometric Society].

Shankar Prinja, Nidhi Gupta, and Ramesh Verma. Censoring in Clinical Trials: Review of Survival Analysis Techniques. *Indian Journal of Community Medicine : Official Publication of Indian Association of Preventive & Social Medicine*, 35(2):217–221, April 2010. ISSN 0970-0218. doi: 10.4103/0970-0218.66859. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2940174/.

Richard Arratia, Larry Goldstein, and Fred Kochman. Size bias for one and all. *Probability Surveys*, 16:1–61, 2019. ISSN 1549-5787. doi: 10.1214/13-PS221. URL https://projecteuclid.org/euclid.ps/1546657438. Publisher: The Institute of Mathematical Statistics and the Bernoulli Society.

Pierre-Jérôme Bergeron. *Covariates and length-biased sampling: is there more than meets the eye?* PhD thesis, McGill University Libraries, [Montreal], 2006. URL https://central.bac-lac.gc.ca/.item?id=TC-QMM-102958&op=pdf&app=Library.

Ying Huang and Mei-Cheng Wang. Estimating the Occurrence Rate for Prevalent Survival Data in Competing Risks Models. *Journal of the American Statistical Association*, 90(432):1406–1415, 1995. ISSN 0162-1459. doi: 10.2307/2291532. URL http://www.jstor.org/stable/2291532. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Mei-Cheng Wang. Nonparametric Estimation from Cross-Sectional Survival Data. *Journal of the American Statistical Association*, 86(413):130–143, 1991. ISSN 0162-1459. doi: 10.2307/2289722. URL http://www.jstor.org/stable/2289722. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

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. URL http://arxiv.org/abs/1806.01984.

Pierre Laforgue and Stephan Clémençon. Statistical Learning from Biased Training Samples. *arXiv:1906.12304 [cs, stat]*, September 2019. URL http://arxiv.org/abs/1906.12304. arXiv: 1906.12304.

V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., 1995. ISBN 0-387-94559-8.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*, volume 103 of *Springer Texts in Statistics*. Springer New York, New York, NY, 2013. ISBN 978-1-4614-7137-0 978-1-4614-7138-7. doi: 10.1007/978-1-4614-7138-7. URL http://link.springer.com/10.1007/978-1-4614-7138-7.

V. N. Vapnik. Principles of Risk Minimization for Learning Theory. In J. Moody, S. Hanson, and R. P. Lippmann, editors, *Advances in Neural Information Processing Systems*, volume 4, pages 831–838. Morgan-Kaufmann, 1992. URL https://proceedings.neurips.cc/paper/1991/file/ff4d5fbbafdf976cfdc032e3bde78de5-Paper.pdf.

V. N. Vapnik and A. Ya. Chervonenkis. The method of ordered risk minimization, i. *Avtomatika i Telemekhanika*, 8:21–30, 1974a.

V. N. Vapnik and A. Ya. Chervonenkis. On the method of ordered risk minimization, ii. *Avtomatika i Telemekhanika*, 9:29–39, 1974b.

Masoud Asgharian, Cyr Emile M'Lan, and David B. Wolfson. Length-Biased Sampling With Right Censoring: An Unconditional Approach. *Journal of the American Statistical Association*, 97(457), 2002. ISSN 0162-1459. doi: 10.1198/016214502753479347. URL https://doi.org/10.1198/016214502753479347. Publisher: Taylor & Francis _eprint: https://doi.org/10.1198/016214502753479347.

George Casella and Roger L. Berger. *Statistical inference*. Thomson Learning, Australia ; Pacific Grove, CA, 2nd ed edition, 2002. ISBN 978-0-534-24312-8.

V. N. Vapnik and A. Ya. Chervonenkis. The uniform convergence of frequencies of the appearance of events to their probabilities. *Doklady Akademii Nauk*, 181(4):781–783, 1968. URL http://mi.mathnet.ru/dan34016.

Brian Sauer, M. Alan Brookhart, Jason A. Roy, and Tyler J. VanderWeele. *Covariate Selection*. Agency for Healthcare Research and Quality (US), January 2013. URL https://www.ncbi.nlm.nih.gov/books/NBK126194/. Publication Title: Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide.

Mohammad Ziaul Islam Chowdhury and Tanvir C. Turin. Variable selection strategies and its importance in clinical prediction modelling. *Family Medicine and Community Health*, 8 (1):e000262, February 2020. ISSN 2305-6983, 2009-8774. doi: 10.1136/fmch-2019-000262. URL https://fmch.bmj.com/content/8/1/e000262. Publisher: BMJ Specialist Journals Section: Methodology.

Robin Genuer, Jean-Michel Poggi, and Christine Tuleau-Malot. Variable selection using random forests. *Pattern Recognition Letters*, 31(14):2225–2236, October 2010a. ISSN 0167-8655. doi: 10.1016/j.patrec.2010.03.014. URL https://www.sciencedirect.com/science/article/pii/S0167865510000954.

Hanna Meyer, Christoph Reudenbach, Stephan Wöllauer, and Thomas Nauss. Importance of spatial predictor variable selection in machine learning applications – Moving from data reproduction to spatial prediction. *Ecological Modelling*, 411:108815, November 2019. ISSN 0304-3800. doi: 10.1016/j.ecolmodel.2019.108815. URL https://www.sciencedirect.com/science/article/pii/S0304380019303230.

Ashkan Ertefaie, Masoud Asgharian, and David A. Stephens. Variable Selection in Causal Inference using a Simultaneous Penalization Method. *Journal of Causal Inference*, 6(1), March 2018. ISSN 2193-3685. doi: 10.1515/jci-2017-0010. URL http://www.degruyter.com/document/doi/10.1515/jci-2017-0010/html. Publisher: De Gruyter Section: Journal of Causal Inference.

Feihan Lu and Eva Petkova. A comparative study of variable selection methods in the context of developing psychiatric screening instruments. *Statistics in medicine*, 33 (3):401–421, February 2014. ISSN 0277-6715. doi: 10.1002/sim.5937. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4026268/.

Loann Desboulets. A Review on Variable Selection in Regression Analysis. *Econometrics*, 6(4):45, November 2018. ISSN 2225-1146. doi: 10.3390/econometrics6040045. URL http://www.mdpi.com/2225-1146/6/4/45.

Georg Heinze, Christine Wallisch, and Daniela Dunkler. Variable selection - A review and recommendations for the practicing statistician. *Biometrical Journal*, 60(3):431–449, May 2018. ISSN 03233847. doi: 10.1002/bimj.201700067. URL http://doi.wiley.com/10.1002/bimj.201700067.

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

Gideon Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 03 1978. doi: 10.1214/aos/1176344136. URL https://doi.org/10.1214/aos/1176344136.

E. J. Hannan and B. G. Quinn. The Determination of the Order of an Autoregression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):190–195, 1979. ISSN 0035-9246. URL http://www.jstor.org/stable/2985032. Publisher: [Royal Statistical Society, Wiley].

Sumio Watanabe. Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *Journal of Machine Learning Research*, 11(116):3571–3594, 2010. URL http://jmlr.org/papers/v11/watanabe10a.html.

Sumio Watanabe. A Widely Applicable Bayesian Information Criterion. *Journal of Machine Learning Research*, 14:867–897, March 2013. ISSN 1533-7928. URL https://jmlr.csail.mit.edu/papers/v14/watanabe13a.html.

Gerda Claeskens and Nils Lid Hjort. The Focused Information Criterion. *Journal of the American Statistical Association*, 98(464):900–916, 2003. ISSN 0162-1459, 1537-274X. doi: 10.1198/016214503000000819. URL http://www.tandfonline.com/doi/abs/10.1198/016214503000000819.

Nils Lid Hjort and Gerda Claeskens. Frequentist Model Average Estimators. *Journal of the American Statistical Association*, 98(464):879–899, 2003. ISSN 0162-1459, 1537-274X. doi: 10.1198/016214503000000828. URL http://www.tandfonline.com/doi/abs/10.1198/016214503000000828.

Nils Lid Hjort and Gerda Claeskens. Focused Information Criteria and Model Averaging for the Cox Hazard Regression Model. *Journal of the American Statistical Association*, 101 (476):1449–1464, 2006. ISSN 0162-1459, 1537-274X. doi: 10.1198/016214506000000069. URL https://www.tandfonline.com/doi/full/10.1198/016214506000000069.

Frank Rosenblatt. *Principles of neurodynamics-perceptrons and the theory of brain mechanisms.* Washington, 1962. URL http://hdl.handle.net/2027/mdp.39015039846566.

Albert Novikoff. On convergence proofs for perceptrons. In *Proceedings of the Symposium on Mathematical Theory of Automata*, volume 12, pages 615–622, Polytechnic Institute of Brooklyn, 1962.

Henry J. Kelley. Gradient Theory of Optimal Flight Paths. *ARS Journal*, 30(10):947–954, October 1960. ISSN 1936-9972. doi: 10.2514/8.5282. URL https://arc.aiaa.org/doi/10.2514/8.5282.

Arthur E. Bryson. A gradient method for optimizing multi-stage allocation processes. In *Proceedings of a Harvard symposium on digital computers and their applications : 3-6 April 1961*. Harvard University Press, 1962.

David E. Rumelhart, Geoffrey E. Hinton, and Williams Williams, Ronald J. Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*, volume 1, pages 318–362. MIT Press, 1986a. ISBN 978-0-262-29140-8. URL https://ieeexplore.ieee.org/document/6302929. Conference Name: Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, October 1986b. ISSN 1476-4687. doi: 10.1038/323533a0. URL https://www.nature.com/articles/323533a0.

V. N. Vapnik. Complete Statistical Theory of Learning. *Automation and Remote Control*, 80(11):1949–1975, 2019. ISSN 1608-3032. doi: 10.1134/S000511791911002X. URL https://doi.org/10.1134/S000511791911002X.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3): 273–297, September 1995. ISSN 0885-6125, 1573-0565. doi: 10.1007/BF00994018. URL http://link.springer.com/10.1007/BF00994018.

R. Bellman, Rand Corporation, and Karreman Mathematics Research Collection. *Dynamic Programming*. Rand Corporation research study. Princeton University Press, 1957. ISBN 9780691079516. URL https://books.google.it/books?id=wdtoPwAACAAJ.

V. N. Vapnik and A. Ya. Chervonenkis. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. *Theory of Probability and Its Applications*, 16(2):264–280, 1971. ISSN 0040-585X, 1095-7219. doi: 10.1137/1116025. URL http://epubs.siam.org/doi/10.1137/1116025.

V. N. Vapnik and A. R. Stepanyuk. Nonparametric Methods for Restoring the Probability Densities. *Avtomatika i Telemekhanika*, 8:38–52, 1978. URL http://mi.mathnet.ru/at9792. Translation: Autom. Remote Control, 39:8 (1979), 1127–1140.

V. N. Vapnik and A. Ya. Chervonenkis. Necessary and Sufficient Conditions for the Uniform Convergence of Means to their Expectations. *Theory of Probability & Its Applications*, 26(3):532–553, 1981. ISSN 0040-585X, 1095-7219. doi: 10.1137/1126059. URL http://epubs.siam.org/doi/10.1137/1126059.

V. N. Vapnik and A. Ya. Chervonenkis. The necessary and sufficient conditions for consistency of the method of empirical risk minimization. In *Yearbook of the Academy of Sciences of the USSR on Recognition, Classification, and Forcasting*, volume 2, pages 217–249, Nauka, Moscow pp. 207–249. Academy of Sciences of the USSR, 1989. URL https://scinapse.io/papers/171746943. English translation: (1991), The necessary and sufficient conditions for consistency of the method of empirical risk minimization, *Pattern Recognition and Image Analysis* 1 (3), pp. 284-305.

V.N. Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5):988–999, September 1999. ISSN 1941-0093. doi: 10.1109/72.788640.

V. N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer Science & Business Media, September 2006. ISBN 978-0-387-34239-9. (Reprint of 1982 Edition).

Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

Frank Rosenblatt. The perceptron—a perceiving and recognizing automaton, 1957.

Frank Rosenblatt. The perceptron — a theory of statistical separability in cognitive systems, January 1958a.

F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958b. ISSN 1939-1471(Electronic),0033-295X(Print). doi: 10.1037/h0042519.

Valery Glivenko. Sulla determinazione empirica delle leggi di probabilita. *Gion. Ist. Ital. Attauri.*, 4:92–99, 1933.

Francesco Paolo Cantelli. Sulla determinazione empirica delle leggi di probabilita. *Giorn. Ist. Ital. Attuari*, 4(421-424), 1933.

Karl Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, July 1900. ISSN 1941-5982, 1941-5990. doi: 10.1080/14786440009463897. URL https://www.tandfonline.com/doi/full/10.1080/14786440009463897.

Karl Pearson. On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is Such that it Can be Reasonably Supposed

to have Arisen from Random Sampling. In Samuel Kotz and Norman L. Johnson, editors, *Breakthroughs in Statistics: Methodology and Distribution*, Springer Series in Statistics, pages 11–28. Springer, New York, NY, 1992. ISBN 978-1-4612-4380-9. doi: 10.1007/978-1-4612-4380-9_2. URL https://doi.org/10.1007/978-1-4612-4380-9_2.

R. A. Fisher. A mathematical Examination of the Methods of determining the Accuracy of Observation by the Mean Error, and by the Mean Square Error. *Monthly Notices of the Royal Astronomical Society*, 80(8):758–770, June 1920. ISSN 0035-8711. doi: 10.1093/mnras/80.8.758. URL https://doi.org/10.1093/mnras/80.8.758.

R. A. Fisher. The Goodness of Fit of Regression Formulae, and the Distribution of Regression Coefficients. *Journal of the Royal Statistical Society*, 85(4):597–612, 1922. ISSN 0952-8385. doi: 10.2307/2341124. URL http://www.jstor.org/stable/2341124. Publisher: [Wiley, Royal Statistical Society].

R. A. Fisher. Theory of Statistical Estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 22(5):700–725, July 1925. ISSN 1469-8064, 0305-0041. doi: 10.1017/S0305004100009580. URL https://www.cambridge.org/core/journals/mathematical-proceedings-of-the-cambridge-philosophical-society/article/abs/theory-of-statistical-estimation/7A05FB68C83B36C0E91D42C76AB177D4. Publisher: Cambridge University Press.

R. A. Fisher. Statistical Methods for Research Workers. In Samuel Kotz and Norman L. Johnson, editors, *Breakthroughs in Statistics: Methodology and Distribution*, Springer Series in Statistics, pages 66–70. Springer, New York, NY, 1992. ISBN 978-1-4612-4380-9. doi: 10.1007/978-1-4612-4380-9_6. URL https://doi.org/10.1007/978-1-4612-4380-9_6.

Ronald Aylmer Fisher. *Statistical methods for research workers*. Oliver and Boyd, Edinburgh, 14th ed., revised and enlarged edition, 1970. ISBN 978-0-05-002170-5.

J. Neyman and E. S. Pearson. On the Problem of the Most Efficient Tests of Statistical Hypotheses. In Samuel Kotz and Norman L. Johnson, editors, *Breakthroughs in Statistics: Foundations and Basic Theory*, Springer Series in Statistics, pages 73–108. Springer, New York, NY, 1992. ISBN 978-1-4612-0919-5. doi: 10.1007/978-1-4612-0919-5_6. URL https://doi.org/10.1007/978-1-4612-0919-5_6.

R. A. Fisher. Inverse probability and the use of Likelihood. *Mathematical Proceedings of the Cambridge Philosophical Society*, 28(3):257–261, July 1932. ISSN 1469-8064, 0305-0041. doi: 10.1017/S0305004100010094. URL https://www.cambridge.org/core/journals/mathematical-proceedings-of-the-

cambridge-philosophical-society/article/abs/inverse-probability-and-the-use-of-likelihood/B3E94B37CA29899A79FE5C0CD0E6A6DD. Publisher: Cambridge University Press.

Samuel S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938.

A.N. Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4:83–91, 1933.

N. V. Smirnov. Sur les écarts de la courbe de distribution empirique. *Matematicheskii Sbornik*, 6(48)(1):3–26, 1939. URL http://mi.mathnet.ru/eng/msb/v48/i1/p3. (in Russian), Publisher: Russian Academy of Sciences, Steklov Mathematical Institute.

John W Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, pages 448–485, 1960.

John W. Tukey. The Future of Data Analysis. *The Annals of Mathematical Statistics*, 33(1):1–67, 1962. ISSN 0003-4851. doi: doi:10.1214/aoms/1177704711. URL http://www.jstor.org/stable/2237638. Publisher: Institute of Mathematical Statistics Permanent link to this document: https://projecteuclid.org/euclid.aoms/1177704711 Mathematical Reviews number (MathSciNet): MR133937 Zentralblatt MATH identifier: 0107.36401.

John Wilder Tukey. *Exploratory data analysis*. Addison-Wesley series in behavioral science. Addison-Wesley Pub. Co, Reading, Mass, 1977. ISBN 978-0-201-07616-5.

Peter J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(1): 73–101, 03 1964. doi: 10.1214/aoms/1177703732. URL https://doi.org/10.1214/aoms/1177703732.

F.R. Hampel. *Contributions to the theory of robust estimation*. University of California, 1968. URL https://books.google.ca/books?id=Ul0ZAQAAIAAJ.

J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972. ISSN 00359238. URL http://www.jstor.org/stable/2344614.

Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, September 2008. ISSN 1932-6157, 1941-7330. doi: 10.1214/08-AOAS169. URL http://projecteuclid.org/journals/annals-of-applied-statistics/volume-

2/issue-3/Random-survival-forests/10.1214/08-AOAS169.full. Publisher: Institute of Mathematical Statistics.

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, December 2008. doi: 10.1109/ICDM.2008.50. ISSN: 2374-8486.

Peter B. Snow, Deborah S. Smith, and William J. Catalona. Artificial Neural Networks in the Diagnosis and Prognosis of Prostate Cancer: A Pilot Study. *The Journal of Urology*, 152 (5, Part 2):1923–1926, November 1994. ISSN 0022-5347. doi: 10.1016/S0022-5347(17)32416-3. URL https://www.sciencedirect.com/science/article/pii/S0022534717324163.

David Faraggi and Richard Simon. A neural network model for survival data. *Statistics in Medicine*, 14(1):73–82, 1995. ISSN 1097-0258. doi: https://doi.org/10.1002/sim.4780140108. URL http://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780140108. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.4780140108.

Blaz Zupan, Janez Demsar, Michael W Kattan, J. Robert Beck, and I Bratko. Machine learning for survival analysis: a case study on recurrence of prostate cancer. *Artificial Intelligence in Medicine*, 20(1):59–75, September 2000. ISSN 0933-3657. doi: 10.1016/S0933-3657(00)00053-1. URL https://www.sciencedirect.com/science/article/pii/S0933365700000531.

José M. Jerez-Aragonés, José A. Gómez-Ruiz, Gonzalo Ramos-Jiménez, José Muñoz-Pérez, and Emilio Alba-Conejo. A combined neural network and decision trees model for prognosis of breast cancer relapse. *Artificial Intelligence in Medicine*, 27(1):45–63, January 2003. ISSN 0933-3657. doi: 10.1016/S0933-3657(02)00086-6. URL https://www.sciencedirect.com/science/article/pii/S0933365702000866.

H.M. Bøvelstad, S. Nygå rd, H.L. Størvold, M. Aldrin, Ø. Borgan, A. Frigessi, and O.C. Lingjærde. Predicting survival from microarray data—a comparative study. *Bioinformatics*, 23(16):2080–2087, August 2007. ISSN 1367-4803. doi: 10.1093/bioinformatics/btm305. URL https://doi.org/10.1093/bioinformatics/btm305.

Daniela M Witten and Robert Tibshirani. Survival analysis with high-dimensional covariates. *Statistical methods in medical research*, 19(1):29–51, February 2010. ISSN 0962-2802. doi: 10.1177/0962280209105024. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4806549/.

Andrew J. Steele, Spiros C. Denaxas, Anoop D. Shah, Harry Hemingway, and Nicholas M. Luscombe. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. *PLOS ONE*, 13(8):e0202344, August 2018. ISSN 1932-6203. doi: 10.1371/journal.pone.0202344. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0202344. Publisher: Public Library of Science.

Laura Macías-García, María Martínez-Ballesteros, José María Luna-Romera, José M. García-Heredia, Jorge García-Gutiérrez, and José C. Riquelme-Santos. Autoencoded DNA methylation data to predict breast cancer recurrence: Machine learning models and gene-weight significance. *Artificial Intelligence in Medicine*, 110:101976, November 2020. ISSN 0933-3657. doi: 10.1016/j.artmed.2020.101976. URL https://www.sciencedirect.com/science/article/pii/S0933365720312410.

Annette Spooner, Emily Chen, Arcot Sowmya, Perminder Sachdev, Nicole A. Kochan, Julian Trollor, and Henry Brodaty. A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Scientific Reports*, 10(1):20410, November 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-77220-w. URL http://www.nature.com/articles/s41598-020-77220-w. Number: 1 Publisher: Nature Publishing Group.

Constantinos Daskalakis, Themis Gouleakis, Christos Tzamos, and Manolis Zampetakis. Computationally and Statistically Efficient Truncated Regression. In *Conference on Learning Theory*, pages 955–960. PMLR, June 2019. URL http://proceedings.mlr.press/v99/daskalakis19a.html. ISSN: 2640-3498.

Constantinos Daskalakis, Dhruv Rohatgi, and Emmanouil Zampetakis. Truncated Linear Regression in High Dimensions. *Advances in Neural Information Processing Systems*, 33:10338–10347, 2020. URL https://papers.nips.cc/paper/2020/hash/751f6b6b02bf39c41025f3bcfd9948ad-Abstract.html.

E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18):2529–2545, September 1999. ISSN 0277-6715. doi: 10.1002/(sici)1097-0258(19990915/30)18:17/18<2529::aid-sim274>3.0.co;2-5.

Patrick J. Heagerty, Thomas Lumley, and Margaret S. Pepe. Time-Dependent ROC Curves for Censored Survival Data and a Diagnostic Marker. *Biometrics*, 56(2):337–344, 2000. ISSN 1541-0420. doi: https://doi.org/10.1111/j.0006-341X.2000.00337.x. URL

https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0006-341X.2000.00337.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.0006-341X.2000.00337.x.

Danh V. Nguyen and David M. Rocke. Partial least squares proportional hazard regression for application to DNA microarray survival data. *Bioinformatics*, 18(12):1625–1632, December 2002. ISSN 1367-4803. doi: 10.1093/bioinformatics/18.12.1625. URL https://doi.org/10.1093/bioinformatics/18.12.1625.

Torsten Hothorn, Berthold Lausen, Axel Benner, and Martin Radespiel-Tröger. Bagging survival trees. *Statistics in Medicine*, 23(1):77–91, 2004. ISSN 1097-0258. doi: https://doi.org/10.1002/sim.1593. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.1593. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.1593.

Hongzhe Li and Jiang Gui. Partial Cox regression analysis for high-dimensional microarray gene expression data. *Bioinformatics*, 20(suppl_1):i208–i215, August 2004. ISSN 1367-4803. doi: 10.1093/bioinformatics/bth900. URL https://doi.org/10.1093/bioinformatics/bth900.

Lexin Li and Hongzhe Li. Dimension reduction methods for microarrays with application to censored survival data. *Bioinformatics*, 20(18):3406–3412, December 2004. ISSN 1367-4803. doi: 10.1093/bioinformatics/bth415. URL https://doi.org/10.1093/bioinformatics/bth415.

Pannagadatta K. Shivaswamy, Wei Chu, and Martin Jansche. A Support Vector Approach to Censored Targets. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 655–660, Omaha, NE, USA, October 2007. IEEE. ISBN 978-0-7695-3018-5. doi: 10.1109/ICDM.2007.93. URL http://ieeexplore.ieee.org/document/4470306/.

Martin Schumacher, Harald Binder, and Thomas Gerds. Assessment of survival prediction models based on microarray data. *Bioinformatics*, 23(14):1768–1774, July 2007. ISSN 1367-4803. doi: 10.1093/bioinformatics/btm232. URL https://doi.org/10.1093/bioinformatics/btm232.

R. Bittern, A. Cuschieri, S. D. Dolgobrodov, R. Marshall, P. Moore, and R. J. C. Steele. An artificial neural network for analysing the survival of patients with colorectal cancer. In *ESANN 2005 Proceedings - 13th European Symposium on Artificial Neural Networks*, pages 103–108, 2007. URL https://discovery.dundee.ac.uk/en/publications/an-artificial-neural-network-for-analysing-the-survival-of-patien.

Yair Goldberg and Michael R. Kosorok. Q-learning with censored data. *The Annals of Statistics*, 40(1):529–560, February 2012. ISSN 0090-5364, 2168-8966. doi: 10.1214/12-AOS968. URL http://projecteuclid.org/journals/annals-of-statistics/volume-40/issue-1/Q-learning-with-censored-data/10.1214/12-AOS968.full. Publisher: Institute of Mathematical Statistics.

Harald Binder. Coxboost: Cox models by likelihood based boosting for a single survival endpoint or competing risks. *R package version*, 1:413–421, 2013.

Badri Padhukasahasram, Chandan Reddy, Yan li, and David Lanfear. Joint Impact of Clinical and Behavioral Variables on the Risk of Unplanned Readmission and Death after a Heart Failure Hospitalization. *PLOS ONE*, 10:e0129553, June 2015. doi: 10.1371/journal.pone.0129553.

Rajesh Ranganath, Adler Perotte, Noémie Elhadad, and David Blei. Deep Survival Analysis. *arXiv:1608.02158 [cs, stat]*, September 2016. URL http://arxiv.org/abs/1608.02158. arXiv: 1608.02158.

Egil Martinsson. Wtte-rnn : Weibull time to event recurrent neural network. Master's thesis, Chalmers University Of Technology, 2016.

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. ISSN 1532-0464. doi: 10.1016/j.jbi.2016.03.009. URL http://www.sciencedirect.com/science/article/pii/S1532046416000496.

Seungyeoun Lee and Heeju Lim. Review of statistical methods for survival analysis using genomic data. *Genomics & Informatics*, 17(4), December 2019. ISSN 1598-866X. doi: 10.5808/GI.2019.17.4.e41. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6944043/.

Mohammadreza Nemati, Jamal Ansary, and Nazafarin Nemati. Machine-Learning Approaches in COVID-19 Survival Analysis and Discharge-Time Likelihood Prediction Using Clinical Data. *Patterns (New York, N.y.)*, 1(5):100074, August 2020. ISSN 2666-3899. doi: 10.1016/j.patter.2020.100074. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7334917/.

Michele De Laurentiis and Peter M. Ravdin. A technique for using neural network analysis to perform survival analysis of censored data. *Cancer Letters*, 77(2-3):127–

138, March 1994. ISSN 03043835. doi: 10.1016/0304-3835(94)90095-7. URL https://linkinghub.elsevier.com/retrieve/pii/0304383594900957.

Knut Liestøl, Per Kragh Andersen, and Ulrich Andersen. Survival analysis and neural nets. *Statistics in Medicine*, 13(12):1189–1200, 1994. ISSN 1097-0258. doi: https://doi.org/10.1002/sim.4780131202. URL http://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780131202. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.4780131202.

Elia Biganzoli, Patrizia Boracchi, Luigi Mariani, and Ettore Marubini. Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Statistics in medicine*, 17(10):1169–1186, 1998.

Brian D. Ripley and Ruth M. Ripley. Neural networks as statistical methods in survival analysis. In Richard Dybowski and Vanya Gant, editors, *Clinical Applications of Artificial Neural Networks*, pages 237–255. Cambridge University Press, Cambridge, 2001. ISBN 978-0-521-66271-0. doi: 10.1017/CBO9780511543494.011. URL https://www.cambridge.org/core/books/clinical-applications-of-artificial-neural-networks/neural-networks-as-statistical-methods-in-survival-analysis/6AC01B644586FE2EF1D34B6A59CC183E.

Rashmi Joshi and Colin Reeves. Beyond the cox model: artificial neural networks for survival analysis part ii. In *Proceedings of the eighteenth international conference on systems engineering*, pages 179–184, 2006.

Chih-Lin Chi, W. Nick Street, and William H. Wolberg. Application of Artificial Neural Network-Based Survival Analysis on Two Breast Cancer Datasets. *AMIA Annual Symposium Proceedings*, 2007:130–134, 2007. ISSN 1942-597X. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2813661/.

Julio Montes-Torres, José Luis Subirats, Nuria Ribelles, Daniel Urda, Leonardo Franco, Emilio Alba, and José Manuel Jerez. Advanced Online Survival Analysis Tool for Predictive Modelling in Clinical Data Science. *PLOS ONE*, 11(8):e0161135, August 2016. ISSN 1932-6203. doi: 10.1371/journal.pone.0161135. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0161135. Publisher: Public Library of Science.

Hamid Nilsaz-Dezfouli, Mohd Rizam Abu-Bakar, Jayanthi Arasan, Mohd Bakri Adam, and Mohamad Amin Pourhoseingholi. Improving Gastric Cancer Outcome Prediction Using Single Time-Point Artificial Neural Network Models. *Cancer Informatics*, 16:

1176935116686062, January 2017. ISSN 1176-9351. doi: 10.1177/1176935116686062. URL https://doi.org/10.1177/1176935116686062. Publisher: SAGE Publications Ltd STM.

Travers Ching, Xun Zhu, and Lana X. Garmire. Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. *PLOS Computational Biology*, 14(4):e1006076, April 2018. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1006076. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006076. Publisher: Public Library of Science.

Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-Event Prediction with Neural Networks and Cox Regression. *Journal of Machine Learning Research*, 20(129):1–30, 2019. ISSN 1533-7928. URL http://jmlr.org/papers/v20/18-424.html.

Yoshua Bengio. An AI Pioneer Wants His Algorithms to Understand the "Why", August 2019. URL https://www.wired.com/story/ai-pioneer-algorithms-understand-why/.

Noor Tilman. Prediction models for survival data with machine learning: an application to soft tissue sarcoma cohort. Master's thesis, Universiteit Leiden, Leiden, Netherlands, March 2020.

James W Hughes and Elizabeth Savoca. Accounting for censoring in duration data: An application to estimating the effect of legal reforms on the duration of medical malpractice disputes. *Journal of Applied Statistics*, 26(2):219–228, 1999.

S. D. Wicksell. The corpuscle problem: A mathematical study of a biometric problem. *Biometrika*, 17(1/2):84–99, 1925. ISSN 00063444. URL http://www.jstor.org/stable/2332027.

S. D. Wicksell. The corpuscle problem: Second memoir: Case of ellipsoidal corpuscles. *Biometrika*, 18(1/2):151–172, 1926. ISSN 00063444. URL http://www.jstor.org/stable/2332500.

R. A. Fisher. The Effect of Methods of Ascertainment Upon the Estimation of Frequencies. *Annals of Eugenics*, 6(1):13–25, 1934. ISSN 2050-1439. doi: 10.1111/j.1469-1809.1934.tb02105.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-1809.1934.tb02105.x.

Jerzy Neyman. Statistics–servant of all sciences. *Science*, 122(3166):401–406, 1955. ISSN 00368075, 10959203. URL http://www.jstor.org/stable/1751553.

JA McFadden. On the lengths of intervals in a stationary point process. *Journal of the Royal Statistical Society: Series B (Methodological)*, 24(2):364–382, 1962.

Saul Blumenthal. Proportional sampling in life length studies. *Technometrics*, 9(2):205–218, 1967.

David Roxbee Cox. Some sampling problems in technology. *New Developments in Survey Sampling*, pages 506–527, 1969.

P L Goldsmith. The calculation of true particle size distributions from the sizes observed in a thin slice. *British Journal of Applied Physics*, 18(6):813–830, jun 1967. doi: 10.1088/0508-3443/18/6/317. URL https://doi.org/10.1088%2F0508-3443%2F18%2F6%2F317.

Harry Smith, Norman Lloyd Johnson, University of North Carolina at Chapel Hill, and University of North Carolina at Chapel Hill, editors. *New Developments in Survey Sampling*. Wiley-Interscience, New York, 1969. ISBN 978-0-471-44487-9. Meeting Name: Symposium on the Foundations of Survey Sampling; OCLC: 33606.

Polly Feigl and Marvin Zelen. Estimation of exponential survival probabilities with concomitant information. *Biometrics*, 21(4):826–838, 1965. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2528247.

M. Zelen. Play the winner rule and the controlled clinical trial. *Journal of the American Statistical Association*, 64(325):131–146, 1969. doi: 10.1080/01621459.1969.10500959. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1969.10500959.

Marvin Zelen. Keynote address on biostatistics and data retrieval. *Cancer chemotherapy reports. Part 3*, 4(2):31, 1973.

Marvin Zelen. A new design for randomized clinical trials. *New England Journal of Medicine*, 300(22):1242–1245, 1979. doi: 10.1056/NEJM197905313002203. URL https://doi.org/10.1056/NEJM197905313002203. PMID: 431682.

Ori Davidov and Marvin Zelen. Referent sampling, family history and relative risk: the role of length-biased sampling. *Biostatistics*, 2(2):173–181, 06 2001. ISSN 1465-4644. doi: 10.1093/biostatistics/2.2.173. URL https://doi.org/10.1093/biostatistics/2.2.173.

Sujuan Gao and Siu L Hui. Estimating the incidence of dementia from two-phase sampling with non-ignorable missing data. *Statistics in medicine*, 19(11-12):1545–1554, 2000.

Yaakov Stern, Min-Xing Tang, Marilyn S. Albert, Jason Brandt, Diane M. Jacobs, Karen Bell, Karen Marder, Mary Sano, Devangere Devanand, Steven M. Albert, Frederick Bylsma, and Wei-Yann Tsai. Predicting Time to Nursing Home Care and Death in Individuals With Alzheimer Disease. *JAMA*, 277(10):806–812, 03 1997. ISSN 0098-7484. doi: 10.1001/jama.1997.03540340040030. URL https://doi.org/10.1001/jama.1997.03540340040030.

Richard Simon. Length-biased sampling in etiologic studies. *American Journal of Epidemiology*, 111(4):444–452, April 1980. ISSN 0002-9262. doi: 10.1093/oxfordjournals.aje.a112920. URL https://academic.oup.com/aje/article/111/4/444/82763.

BB Winter and A Földes. A product-limit estimator for use with length-biased data. *Canadian Journal of Statistics*, 16(4):337–355, 1988.

Clifford Nowell, Marc A. Evans, and Lyman McDonald. Length-biased sampling in contingent valuation studies. *Land Economics*, 64(4):367–371, 1988. ISSN 00237639. URL http://www.jstor.org/stable/3146309.

Clifford Nowell and Linda R. Stanley. Length-biased sampling in mall intercept surveys. *Journal of Marketing Research*, 28(4):475–479, 1991. doi: 10.1177/002224379102800409. URL https://doi.org/10.1177/002224379102800409.

J De Uña Álvarez. On efficiency under selection bias caused by truncation. *Unpublished manuscript*, 2001.

Samuel Stanley Wilks. *Mathematical Statistics*. s.n., repr. der ausg. princeton, nj, 1943 edition, 1943. ISBN 978-1-4067-3431-7. OCLC: 555161827.

I. A. Ibragimov and R. Z. Has'minskii. *Statistical Estimation: Asymptotic Theory.*, volume 16 of *Applications of mathematics*. Springer, 1981. ISBN Print: 978-1-4899-0029-6; Online: 978-1-4899-0027-2. URL https://www.springer.com/gp/book/9781489900272. OCLC: 1080425596.

Theodoros Evgeniou, Massimiliano Pontil, and Tomaso Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13(1):1, April 2000. ISSN 1572-9044. doi: 10.1023/A:1018946025316. URL https://doi.org/10.1023/A:1018946025316.

Albert Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*. Other Titles in Applied Mathematics. Society for Industrial and Applied Mathematics, January 2005. ISBN 978-0-89871-572-9. doi: 10.1137/1.9780898717921. URL https://epubs.siam.org/doi/book/10.1137/1.9780898717921.

Tom M. Mitchell. The need for biases in learning generalizations. Technical report, Rutgers University, New Brunswick, NJ, 1980.

Diana Gordon and Marie desJardins. Evaluation and selection of biases in machine learning. *Machine Learning*, 20:5–22, 07 1995. doi: 10.1007/BF00993472.

WO Whitney and CJ Mehlhaff. High-rise syndrome in cats. *Journal of the American Veterinary Medical Association*, 191(11):1399—1403, December 1987. ISSN 0003-1488.

David W. Hosmer, Stanley Lemeshow, and Susanne May. *Applied survival analysis: regression modeling of time-to-event data*. Wiley series in probability and statistics. Wiley-Interscience, Hoboken, N.J, 2nd edition, 2008. ISBN 978-0-471-75499-2. OCLC: ocn154798742.

C. Radhakrishna Rao. On discrete distributions arising out of methods of ascertainment. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 27(2/4):311–324, 1965. ISSN 0581572X. URL http://www.jstor.org/stable/25049375.

C. Radhakrishna Rao. Weighted Distributions Arising Out of Methods of Ascertainment: What Population Does a Sample Represent? In Anthony C. Atkinson and Stephen E. Fienberg, editors, *A Celebration of Statistics*, pages 543–569, New York, NY, 1985. Springer. ISBN 978-1-4613-8560-8. doi: 10.1007/978-1-4613-8560-8_24.

Larry V. Hedges. Modeling Publication Selection Effects in Meta-Analysis. *Statistical Science*, 7(2):246–255, 1992. ISSN 0883-4237. URL http://www.jstor.org/stable/2246311. Publisher: Institute of Mathematical Statistics.

Satish Iyengar and Joel B. Greenhouse. Selection Models and the File Drawer Problem. *Statistical Science*, 3(1):109–117, 1988. ISSN 0883-4237. URL http://www.jstor.org/stable/2245925. Publisher: Institute of Mathematical Statistics.

Ron Brookmeyer and Mitchell H. Gail. Biases in prevalent cohorts. *Biometrics*, 43(4): 739–749, 1987. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2531529.

David Roxbee Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

N. E. Breslow. Analysis of Survival Data under the Proportional Hazards Model. *International Statistical Review / Revue Internationale de Statistique*, 43(1):45–57, 1975. ISSN 0306-7734. doi: 10.2307/1402659. URL http://www.jstor.org/stable/1402659. Publisher: [Wiley, International Statistical Institute (ISI)].

David Schoenfeld. Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika*, 67(1):145–153, 1980. ISSN 00063444. URL http://www.jstor.org/stable/2335327.

David Schoenfeld. Partial residuals for the proportional hazards regression model. *Biometrika*, 69(1):239–241, 1982. ISSN 0006-3444, 1464-3510. doi: 10.1093/biomet/69.1.239. URL https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/69.1.239.

Patricia M. Grambsch and Terry M. Therneau. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3):515–526, 1994. ISSN 0006-3444, 1464-3510. doi: 10.1093/biomet/81.3.515. URL https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/81.3.515.

Jonathan Buckley and Ian James. Linear Regression with Censored Data. *Biometrika*, 66 (3):429–436, 1979. ISSN 0006-3444. doi: 10.2307/2335161. URL http://www.jstor.org/stable/2335161. Publisher: [Oxford University Press, Biometrika Trust].

Y. Vardi. Multiplicative censoring, renewal processes, deconvolution and decreasing density: Nonparametric estimation. *Biometrika*, 76(4):751–761, 1989. ISSN 00063444. URL http://www.jstor.org/stable/2336635.

William Feller. *An Introduction to Probability Theory and Its Applications, Volume 2*, volume 2 of *A Wiley publication in mathematical statistics*. Wiley, New York, 1st edition, 1971.

E. L. Kaplan and Paul Meier. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282):457–481, June 1958. ISSN 0162-1459. doi: 10.1080/01621459.1958.10501452. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1958.10501452.

Lucien Le Cam and Grace Lo Yang. *Asymptotics in Statistics: Some Basic Concepts*. Springer Series in Statistics (SSS). Springer New York : Imprint : Springer, New York, NY, 1st edition, 1990. ISBN 978-1-4612-1166-2 978-1-4612-7030-0 978-1-4684-0377-0. URL https://doi.org/10.1007/978-1-4612-1166-2. OCLC: 840278016 DOl: 10.1007/978-1-4684-0377-0.

E. A. Nadaraya. On Estimating Regression. *Theory of Probability & Its Applications*, 9 (1):141–142, 1964. ISSN 0040-585X, 1095-7219. doi: 10.1137/1109020. URL http://epubs.siam.org/doi/10.1137/1109020.

Geoffrey S. Watson. Smooth Regression Analysis. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 26(4):359–372, 1964. ISSN 0581-572X. URL http://www.jstor.org/stable/25049340. Publisher: Springer.

Min Zhang, Dabao Zhang, and Martin T. Wells. Variable selection for large p small n regression models with incomplete data: Mapping QTL with epistases. *BMC Bioinformatics*, 9(1):251, May 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-251. URL https://doi.org/10.1186/1471-2105-9-251.

N. Akarachantachote, S. Chadcham, and K. Saithanu. CUTOFF THRESHOLD OF VARIABLE IMPORTANCE IN PROJECTION FOR VARIABLE SELECTION. *International Journal of Pure and Applied Mathematics*, 94(3):307–322, 2014. ISSN 1311-8080. URL https://ijpam.eu/contents/2014-94-3/2/index.html. Publisher: Academic Publications, Ltd.

Robin Genuer, Vincent Michel, Evelyn Eger, and Bertrand Thirion. Random forests based feature selection for decoding fmri data. In *Proceedings Compstat*, volume 267, pages 1–8, 2010b.

E. L. Lehmann. *Elements of large-sample theory*. Springer texts in statistics. Springer, New York, 1999. ISBN 978-0-387-98595-4.

E. L. Lehmann. *Theory of point estimation*. Wiley series in probability and mathematical statistics. Wiley, New York, 1983. ISBN 978-0-471-05849-6.

E. L. Lehmann and George Casella. *Theory of point estimation*. Springer Texts in Statistics (STS). Springer-Verlag New York, New York, 2nd ed edition, 1998. ISBN Print: 978-0-387-98502-2; Online: 978-0-387-22728-3. URL https://www.springer.com/gp/book/9780387985022. DOI https://doi-org.proxy3.library.mcgill.ca/10.1007/b98854 Originally published as a monograph.

Colin B. Begg and Robert J. Gray. Methodology for case-control studies with prevalent cases. *Biometrika*, 74(1):191–195, 03 1987. ISSN 0006-3444. doi: 10.1093/biomet/74.1.191. URL https://doi.org/10.1093/biomet/74.1.191.

Fadil Santosa and William W Symes. Linear inversion of band-limited reflection seismograms. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1307–1330, 1986.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 00359246. URL http://www.jstor.org/stable/2346178.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006. doi: 10.1198/016214506000000735. URL https://doi.org/10.1198/016214506000000735.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

Xavier Guyon and Jian-feng Yao. On the Underfitting and Overfitting Sets of Models Chosen by Order Selection Criteria. *Journal of Multivariate Analysis*, 70(2):221–249, August 1999. ISSN 0047-259X. doi: 10.1006/jmva.1999.1828. URL https://www.sciencedirect.com/science/article/pii/S0047259X99918286.

Omidali Aghababaei Jazi. *Semiparametric estimation and variable selection under length-biased sampling with heavy censoring*. PhD Thesis, McGill University, Montreal, 2019. URL https://escholarship.mcgill.ca/concern/theses/q237hx212?locale=en.

Hamparsum Bozdogan and Dominique M. A. Haughton. Informational complexity criteria for regression models. *Computational Statistics & Data Analysis*, 28(1):51–76, July 1998. ISSN 0167-9473. doi: 10.1016/S0167-9473(98)00025-5. URL http://www.sciencedirect.com/science/article/pii/S0167947398000255.

Jun Shao. Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88(422):486–494, 1993. ISSN 01621459. URL http://www.jstor.org/stable/2290328.

Jianqing Fan and Heng Peng. Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961, June 2004. ISSN 0090-5364, 2168-8966. doi: 10.1214/009053604000000256. URL http://projecteuclid.org/journals/annals-of-statistics/volume-32/issue-3/Nonconcave-penalized-likelihood-with-a-diverging-number-of-parameters/10.1214/009053604000000256.full. Publisher: Institute of Mathematical Statistics.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York, New York, NY, 2009.

ISBN 978-0-387-84857-0 978-0-387-84858-7. doi: 10.1007/978-0-387-84858-7. URL http://link.springer.com/10.1007/978-0-387-84858-7.

Anestis Antoniadis and Jianqing Fan. Regularization of Wavelet Approximations. *Journal of the American Statistical Association*, 96(455):939–955, 2001. ISSN 0162-1459. URL http://www.jstor.org/stable/2670237. Publisher: American Statistical Association, Taylor & Francis, Ltd.

David Hilbert. Mathematical problems. *Bulletin of the American Mathematical Society*, 8 (10):437–479, 1902.

David Hilbert. Mathematische probleme. *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*, pages 253–297, 1900. URL http://www.digizeitschriften.de/dms/img/?PID=GDZPPN002498863.

A. G. Vitushkin. On hilbert's thirteenth problem and related questions. *Russian Mathematical Surveys*, 59(1):11–25, feb 2004. doi: 10.1070/rm2004v059n01abeh000698. URL https://doi.org/10.1070%2Frm2004v059n01abeh000698.

George Pólya and Gabor Szegö. Some Properties of Real Functions. In George Pólya and Gabor Szegö, editors, *Problems and Theorems in Analysis I: Series. Integral Calculus. Theory of Functions*, Classics in Mathematics, pages 75–84. Springer, Berlin, Heidelberg, 1998. ISBN 978-3-642-61983-0. URL https://doi.org/10.1007/978-3-642-61983-0_17.

Vladimir I. Arnold. *Vladimir I. Arnold - Collected Works: Representations of Functions, Celestial Mechanics, and KAM Theory 1957-1965*, volume 1. Springer-Verlag, Berlin Heidelberg, 2009a. ISBN 978-3-642-01741-4. doi: 10.1007/978-3-642-01742-1. URL https://www.springer.com/gp/book/9783642017414.

A. G. Vitushkin. On multidimensional variations. *Gostekhizdat, Moscow*, 1955.

A. N. Kolmogorov. On the representation of continuous functions of several variables by superpositions of continuous functions of a smaller number of variables. *American Mathematical Society Translations: Series 2*, 17:369–373, 1961. doi: http://dx.doi.org/10.1090/trans2/017.

V. I. Arnold. On the representation of functions of several variables as a superposition of functions of a smaller number of variables. In *Vladimir I. Arnold - Collected Works: Representations of Functions, Celestial Mechanics, and KAM Theory 1957-1965*, pages 25–45. Springer-Verlag, Berlin Heidelberg, 2009b. ISBN 978-3-642-01741-4. doi: 10.1007/978-

3-642-01742-1. URL https://www.springer.com/gp/book/9783642017414. (Original work published 1958, in Russian).

V. I. Arnold. On functions of three variables. *Doklady Akademii Nauk*, 114(4):679–681, 1957. (Russian).

A. N. Kolmogorov. On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition. In Vladimir M. Tikhomirov, editor, *Selected Works of A. N. Kolmogorov: Volume I: Mathematics and Mechanics*, pages 383–387. Springer Science & Business Media, 1991. ISBN 978-90-277-2796-1.

Alexander Ostrowski. Über Dirichletsche Reihen und algebraische Differentialgleichungen. *Mathematische Zeitschrift*, 8(3):241–298, September 1920. ISSN 1432-1823. doi: 10.1007/BF01206530. URL https://doi.org/10.1007/BF01206530.

A. G. Vitushkin and G. M. Khenkin. Linear superpositions of functions. *Russian Mathematical Surveys*, 22(1):77–125, February 1967. ISSN 0036-0279, 1468-4829. doi: 10.1070/RM1967v022n01ABEH001204. URL http://stacks.iop.org/0036-0279/22/i=1/a=R03?key=crossref.64d409a892119b2ecf27a3a8a047e1d4.

A. G. Vitushkin. On Hilbert's Thirteenth Problem. *Dokl. Akad. Nauk SSSR, Novaya. Seriya.*, 95(4):701–704, 1954. ISSN 0002-3264.

V. N. Tikhomirov. Kolmogorov's work on ε-entropy of functional classes and the superposition of functions. *Russian Mathematical Surveys*, 18(5):51–87, oct 1963. doi: 10.1070/rm1963v018n05abeh004132. URL https://doi.org/10.1070%2Frm1963v018n05abeh004132.

A. N. Kolmogorov. Estimates of the minimal number of elements of ε-nets in various functional classes and their application to the question of representability of functions of several variables by superpositions of functions of fewer variables. *Uspekhi Mat. Nauk*, 10 (1):192–193, 1955. (Also in: Dokl. Akad. Nauk SSSR 101 (1955), 192–194).

Anatoliy Georgievich Vitushkin. A proof of the existence of analytic functions of several variables not representable by linear superpositions of continuously differentiable functions of fewer variables. *Doklady Akademii Nauk*, 156(6):1258–1261, 1964. URL http://mi.mathnet.ru/dan29742.

Gennadi Markovich Khenkin. Linear superpositions of continuously differentiable functions. *Doklady Akademii Nauk*, 157(2):288–290, 1964. (in Russian).

A. A. Milyutin. Isomorphism of the spaces of continuous functions on compacta with the cardinality of the continuum. *Theory of Functions, Functional Analysis and their Applications*, pages 150–156, 1966. URL http://dspace.univer.kharkov.ua/handle/123456789/189. (in Russian).

G. G. Lorentz. Metric entropy, widths, and superpositions of functions. *The American Mathematical Monthly*, 69(6):469–485, 1962. ISSN 00029890, 19300972. URL http://www.jstor.org/stable/2311185.

David A. Sprecher. On the structure of continuous functions of several variables. *Transactions of the American Mathematical Society*, 115:340–355, 1965. ISSN 0002-9947. doi: 10.2307/1994273. URL https://www.jstor.org/stable/1994273. Publisher: American Mathematical Society.

David A. Sprecher. On the structure of representations of continuous functions of several variables as finite sums of continuous functions of one variable. *Proceedings of the American Mathematical Society*, 17(1):98–105, 1966. ISSN 0002-9939. doi: 10.2307/2035068. URL https://www.jstor.org/stable/2035068.

David A. Sprecher. An improvement in the superposition theorem of Kolmogorov. *Journal of Mathematical Analysis and Applications*, 38(1):208–213, April 1972. ISSN 0022-247X. doi: 10.1016/0022-247X(72)90129-1. URL http://www.sciencedirect.com/science/article/pii/0022247X72901291.

Phillip A. Ostrand. Dimension of metric spaces and hilbert's problem 13. *Bulletin of the American Mathematical Society*, 71(4):619–622, 1965. ISSN 0002-9904, 1936-881X. doi: 10.1090/S0002-9904-1965-11363-5. URL https://www.ams.org/bull/1965-71-04/S0002-9904-1965-11363-5/.

B. L. Fridman. An improvement in the smoothness of the functions in a. n. kolmogorov's theorem on superpositions. *Doklady Akademii Nauk*, 177(5):1019–1022, 1967. (in Russian).

David A. Sprecher. A universal mapping for kolmogorov's superposition theorem. *Neural Networks*, 6(8):1089–1094, January 1993. ISSN 0893-6080. doi: 10.1016/S0893-6080(09)80020-8. URL http://www.sciencedirect.com/science/article/pii/S0893608009800208.

Albert B. Novikoff. On convergence proofs for perceptrons. Technical report, Stanford Research Institute, Menlo Park, CA, 1963.

Stuart Dreyfus. The numerical solution of variational problems. *Journal of Mathematical Analysis and Applications*, 5(1):30–45, August 1962. ISSN 0022247X. doi: 10.1016/0022-247X(62)90004-5. URL https://linkinghub.elsevier.com/retrieve/pii/0022247X62900045.

Seppo Linnainmaa. Taylor expansion of the accumulated rounding error. *BIT*, 16(2): 146–160, June 1976. ISSN 0006-3835, 1572-9125. doi: 10.1007/BF01931367. URL http://link.springer.com/10.1007/BF01931367.

D. B. Parker. Learning-Logic: Casting the Cortex of the Human Brain in Silicon. Technical Report Tr-47, Center for Computational Research in Economics and Management Science, MIT Cambridge, MA., 1985.

Paul J. Werbos. *The roots of backpropagation: from ordered derivatives to neural networks and political forecasting*. Adaptive and learning systems for signal processing, communications, and control. Wiley, New York, 1994. ISBN 978-0-471-59897-8.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. Adaptive computation and machine learning. The MIT Press, Cambridge, Massachusetts, 2016. ISBN 978-0-262-03561-3.

Seppo Linnainmaa. The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors. Master's thesis, University of Helsinki, Helsinki, Finland, 1970.

Robert Hecht-Nielsen. Kolmogorov's mapping neural network existence theorem. In *Proceedings of the IEEE International Conference on Neural Networks*, volume 3, pages 11–14, New York, NY, 1987. IEEE Press.

Federico Girosi and Tomaso Poggio. Representation Properties of Networks: Kolmogorov's Theorem Is Irrelevant. *Neural Computation*, 1(4):465–469, December 1989. ISSN 0899-7667. doi: 10.1162/neco.1989.1.4.465. URL https://doi.org/10.1162/neco.1989.1.4.465. Publisher: MIT Press.

Vera Kurková. Kolmogorov's theorem is relevant. *Neural Computation*, 3(4):617–622, 1991. doi: 10.1162/neco.1991.3.4.617. URL https://doi.org/10.1162/neco.1991.3.4.617. PMID: 31167327.

Vera Kurková. Kolmogorov's theorem and multilayer neural networks. *Neural Networks*, 5(3):501–506, January 1992. ISSN 0893-6080. doi: 10.1016/0893-6080(92)90012-8. URL http://www.sciencedirect.com/science/article/pii/0893608092900128.

G Cybenko. Continuous Valued Neural Networks with Two Hidden Layers are Sufficient. Technical report, Department of Computer Science, Tufts University, Massachusetts, 1988. OCLC: 123325317.

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, December 1989. ISSN 1435-568X. doi: 10.1007/BF02551274. URL https://doi.org/10.1007/BF02551274.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer Feedforward Networks Are Universal Approximators, 1988. Discussion Paper 88-45. San Diego, CA: Department of Economics, University of California, San Diego.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, January 1989. ISSN 0893-6080. doi: 10.1016/0893-6080(89)90020-8. URL http://www.sciencedirect.com/science/article/pii/0893608089900208.

Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, January 1991. ISSN 0893-6080. doi: 10.1016/0893-6080(91)90009-T. URL http://www.sciencedirect.com/science/article/pii/089360809190009T.

Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, January 1993. ISSN 0893-6080. doi: 10.1016/S0893-6080(05)80131-5. URL http://www.sciencedirect.com/science/article/pii/S0893608005801315.

David A. Sprecher. A Numerical Implementation of Kolmogorov's Superpositions. *Neural Networks*, 9(5):765–772, July 1996. ISSN 0893-6080. doi: 10.1016/0893-6080(95)00081-X. URL http://www.sciencedirect.com/science/article/pii/089360809500081X.

David A. Sprecher. A numerical implementation of kolmogorov's superpositions II. *Neural Networks*, 10(3):447–457, April 1997. ISSN 0893-6080. doi: 10.1016/S0893-6080(96)00073-1. URL http://www.sciencedirect.com/science/article/pii/S0893608096000731.

B. Igelnik and N. Parikh. Kolmogorov's spline network. *IEEE Transactions on Neural Networks*, 14(4):725–733, July 2003. ISSN 1941-0093. doi: 10.1109/TNN.2003.813830.

Jürgen Braun and Michael Griebel. On a constructive proof of kolmogorov's superposition theorem. *Constructive Approximation*, 30:653–675, 2009. doi: https://doi.org/10.1007/s00365-009-9054-2.

Jürgen Braun. *An Application of Kolmogorov's Superposition Theorem to Function Reconstruction in Higher Dimensions*. PhD Thesis, Mathematisch–Naturwissenschaftlichen Fakultät der Rheinischen Friedrich–Wilhelms–Universität Bonn, Bonn, Germany, 2009.

Leila Ait Gougam, Mouloud Tribeche, and Fawzia Mekideche-Chafa. A systematic investigation of a neural network for function approximation. *Neural Networks*, 21(9): 1311–1317, November 2008. ISSN 0893-6080. doi: 10.1016/j.neunet.2008.06.015. URL http://www.sciencedirect.com/science/article/pii/S0893608008001378.

Ronen Eldan and Ohad Shamir. The Power of Depth for Feedforward Neural Networks. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 907–940, Columbia University, New York, New York, USA, June 2016. PMLR. URL http://proceedings.mlr.press/v49/eldan16.html.

Tomaso Poggio, Hrushikesh Mhaskar, Lorenzo Rosasco, Brando Miranda, and Qianli Liao. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: A review. *International Journal of Automation and Computing*, 14(5):503–519, 2017. ISSN 1751-8520. doi: 10.1007/s11633-017-1054-2. URL https://doi.org/10.1007/s11633-017-1054-2.

Dmitry Yarotsky. Optimal approximation of continuous functions by very deep relu networks. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 639–649. PMLR, 06–09 Jul 2018. URL http://proceedings.mlr.press/v75/yarotsky18a.html.

Philipp Petersen and Felix Voigtlaender. Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Networks*, 108:296–330, December 2018. ISSN 0893-6080. doi: 10.1016/j.neunet.2018.08.019. URL http://www.sciencedirect.com/science/article/pii/S0893608018302454.

Bo Liu and Yi Liang. Optimal Function Approximation with Relu Neural Networks. *ArXiv*, 2019. URL https://arxiv.org/abs/1909.03731.

Ilsang Ohn and Yongdai Kim. Smooth Function Approximation by Deep Neural Networks with General Activation Functions. *Entropy*, 21(7), July 2019. doi: 10.3390/e21070627. URL https://www.mdpi.com/1099-4300/21/7/627.

Minshuo Chen, Haoming Jiang, Wenjing Liao, and Tuo Zhao. Efficient Approximation of Deep ReLU Networks for Functions on Low Dimensional Manifolds. In *Advances in Neural Information Processing Systems 32*, pages 8174–8184. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/9028-efficient-approximation-of-deep-relu-networks-for-functions-on-low-dimensional-manifolds.pdf.

Hadrien Montanelli and H. Yang. Error bounds for deep ReLU networks using the Kolmogorov-Arnold superposition theorem. *Neural networks : the official journal of the International Neural Network Society*, 2020. doi: 10.1016/j.neunet.2019.12.013.

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. doi: 10.1109/ICDM.2008.50.

Yair Goldberg and Michael R. Kosorok. Support vector regression for right censored data. *Electron. J. Statist.*, 11(1):532–569, 2017. doi: 10.1214/17-EJS1231. URL https://doi.org/10.1214/17-EJS1231.

Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018. ISBN 978-1-5416-1698-1. URL http://public.ebookcentral.proquest.com/choice/publicfullrecord.aspx?p=5368826. OCLC: 1053546137.

Bernhard Schölkopf. Causality for Machine Learning. *arXiv:1911.10500 [cs, stat]*, December 2019. URL http://arxiv.org/abs/1911.10500. arXiv: 1911.10500.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning series. The MIT Press, Cambridge, Massachusetts and London, England, 2017. ISBN 978-0-262-03731-0. URL https://mitpress.mit.edu/books/elements-causal-inference.