From the Unit of Biostatistics,
Institute of Environmental Medicine,
Karolinska Institutet, Stockholm, Sweden

MATHEMATICAL PROGRAMMING FOR OPTIMAL PROBABILITY WEIGHTING

Michele Santacatterina

Stockholm 2018

All published papers reproduced with permission
Published by Karolinska Institutet
Printed by E-Print AB 2018

Typeset by the author using $\LaTeX\,2_\varepsilon$

ISBN 978-91-7831-005-0

MATHEMATICAL PROGRAMMING FOR OPTIMAL PROBABILITY WEIGHTING

THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Michele Santacatterina

Principal supervisor:
Professor Matteo Bottai
Karolinska Institutet
Institute of Environmental Medicine

Co-supervisors:
Professor Rino Bellocco
University of Milano-Bicocca
Department of Statistics and Quantitative Methods

Professor Anna Mia Ekström
Karolinska Institutet
Department of Public Health Sciences

Professor Anders Sonnerbörg
Karolinska Institutet
Department of Medicine

Opponent:
Assistant Professor José Ramon Zubizarreta
Harvard University
Department of Health Care Policy

Examination board:
Professor Timo Koski
Royal Institute of Technology
Department of Mathematics

Senior Lecturer Mark Clements
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Associate Professor Anna Grimby Ekman
University of Gothenburg
Department of Public Health and Community Medicine

Abstract

In spite of the fact that probability weighting is widely used in statistics to correct for unequal sampling, control for confounding, and handle missing data, it has two main limitations. First, statistical inferences may be inefficient in the presence of extreme probability weights. Second, probability weighting-based methods are highly sensitive to model misspecifications. The aim of this Ph.D. thesis work was to develop novel methods, based on mathematical programming techniques, for optimal probability weighting. Specifically, in Paper I, we proposed a method that estimates optimal probability weights, which are obtained as the solution to a constrained optimization problem that minimizes the Euclidean distance from the target (original/design) weights among all sets of weights that satisfy a constraint on the precision of the resulting weighted estimator. In Paper II, we extended optimal probability weights to estimate the causal effect of a time-varying treatment on a survival outcome. Optimal probability weights were obtained as the solution to a constrained optimization problem which constrained the variance of the weights, rather than the standard error of the resulting weighted estimator, as in Paper I. In Paper III, we proposed Kernel Optimal Weighting (KOW), to obtain weights that optimally balance time-dependent confounders while controlling for the precision of the resulting marginal structural model estimate by directly minimizing the error in estimation. This error is expressed as an operator derived from the g-computation formula and KOW minimizes its operator norm with respect to a reproducing kernel Hilbert spaces by solving a quadratic optimization problem. KOW mitigates the effects of possible misspecification of the treatment model by directly balancing covariates and control for precision by penalizing extreme weights. In Paper IV, we evaluated the effect of treatment switch on time to second-line HIV treatment failure using data from the Swedish InfCare HIV registry. This Ph.D. thesis provided methods that will likely help to (1) extend the use of probability weighting in medicine, epidemiology, and economics, (2) extend knowledge on how mathematical programming and machine learning could be used to conduct robust analyses for improved decision-making, and, (3) provide powerful, strong, and robust results to clinicians and policy-makers.

List of publications

I. Michele Santacatterina, and Matteo Bottai
Optimal probability weights for inference with constrained precision
Journal of the American Statistical Association 2017; in press

II. Michele Santacatterina, Rino Bellocco, Anders Sönnerborg, Anna Mia Ekström, and Matteo Bottai
Optimal probability weights for estimating causal effects of time-varying treatments with marginal structural Cox models
Submitted 2018;

III. Michele Santacatterina, and Nathan Kallus
Optimal balancing of time-dependent confounders for marginal structural models
Manuscript 2018;

IV. Amanda Häggblom, Michele Santacatterina, Ujjwal Neogi, Magnus Gisslen, Bo Hejdeman, Leo Flamholc, and Anders Sönnerborg
Effect of therapy switch on time to second-line antiretroviral treatment failure in HIV-infected patients
PloS one 2017; 12(7):e0180140

The articles will be referred to in the text by their Roman numerals, and are reproduced in full at the end of the thesis.

Other publications

- Michele Santacatterina, and Matteo Bottai
  Inferences and conjectures in clinical trials: a systematic review of generalizability of study findings
  Journal of internal medicine 2016; 276(1):123–126

- Niklas Karlsson, Michele Santacatterina, Kerstin Käll, Maria Hägerstrand, Susanne Wallin, Torsten Berglund, and Anna Mia Ekström
  Risk behaviour determinants among people who inject drugs in Stockholm, Sweden over a 10-year period, from 2002 to 2012
  Harm reduction journal 2017; 14(1)

Contents

1 Introduction **1**

2 Background **3**

2.1 Weighting in survey sampling 3

2.1.1 Calibration Estimators 3

2.2 Missing data 4

2.3 Causal Inference 5

2.3.1 Inverse probability of treatment weighting 6

2.3.2 Matching 7

2.3.3 Covariate balancing 8

2.3.4 Causal inference in longitudinal studies 8

2.4 Limitations of probability weighting methods 10

2.4.1 Extreme probability weights 10

2.4.2 Model misspecification 10

2.5 Kernels and Gaussian processes 11

2.5.1 Kernels and reproducing kernel Hilbert spaces 11

2.5.2 Gaussian processes for machine learning 11

2.6 Mathematical Programming 12

3 Aims of the thesis **13**

4 Summary of the studies **14**

4.1 Paper I 14

4.1.1 Additional results 15

4.2 Paper II 15

4.3 Paper III 20

4.4 Paper IV 21

4.4.1 Additional results 21

5 Discussion **23**

6 Future research **24**

References **27**

Acknowledgements **34**

Chapter 1

Introduction

Probability weighting is widely used in statistics to (1) correct for unequal sampling fractions, i.e. sampling weighting (Pfeffermann, 1993); (2) estimate the causal effect of a treatment (or intervention) from observational studies, i.e. inverse probability of treatment weighting (Robins et al., 1994); (3) handle missing data (Little and Rubin, 2014) and (4) generalize the study results from randomized trials (Stuart et al., 2011a).

The key idea of probability weighting is to correct for sample's disproportionalities with respect to a target population of interest by weighting each unit in the sample. For example, when estimating the causal effect of a treatment on an outcome with observational data, the target population of interest is that in which covariates are balanced across treatment groups. Each unit is consequently weighted by the inverse of the probability of being assigned to the treatment conditional on covariates.

It is well known, however, that statistical inference may be inefficient when weights contain outlying values (Rao, 1966; Basu, 2011; Robins et al., 2007, 1995; Scharfstein et al., 1999, among others). In addition, in the context of causal inference, inverse probability of treatment weighting methods are highly sensitive to misspecifications of the treatment model (Kang and Schafer, 2007b; Lefebvre et al., 2008; Cole and Hernán, 2008).

Mathematical programming, also know as optimization, is a branch of mathematics that concerns the theory and development of methods that find extrema of an objective function (Luenberger and Ye, 2015). Examples include running a business in which profit has to be maximized or losses minimized, and the selection of a flight route which minimizes the fuel cost. Mathematical programming is also used in the context of probability weighting. For example, calibration estimators in survey sampling use calibration weights which are obtained by solving a constrained optimization problem (Deville and Särndal, 1992). Optimal matching, used to balance covariates in observational studies, can be reinterpreted as a network flow optimization problem (Rosenbaum, 1989).

In this Ph.D. thesis, we present methods, based on mathematical programming techniques, that (1) provide a set of weights, called optimal probability weights, which are the closest to the target (original/design) weights of interest while controlling the precision of the resulting weighted estimator; (2) extend optimal probability weights to the estimation of the causal effect of a time-varying treatment on a survival outcome; and (3) provide weights that optimally balance time-dependent confounders, thus mitigating the effects of possible misspecification of the treatment model, and control for the precision of the resulting marginal structural model estimate. We apply the proposed methods to the study of the effect of timing of treatment initiation on long-term treatment efficacy in patients infected with human immunodeficiency virus (HIV).

A motivational example

This Ph.D. work was motivated by the abundant use of probability weighting in observational studies to estimate the effect of treatment initiation among people who live with HIV (Hernán et al., 2000, 2001; HIV-Causal Collaboration et al., 2010, 2011; Lodi et al., 2017, among others). For example, Cole and Hernán (2008) evaluated the effect of HIV treatment initiation on the evolution of CD4 cell count by estimating the parameters of the marginal structural model using inverse probability of treatment weighting. Weights were obtained as the inverse of the probability of treatment initiation given a set of baseline covariates. The authors concluded that the effect of HIV treatment initiation on differences in CD4 cell count after one year of treatment was equal to 29 with a 95% confidence interval (CI) equal to (15.8;42.3). The unweighted analysis reported an effect equal to -6 with a 95% CI equal to (-4.4;3.1). The weighted estimate was highly variable, with a confidence interval almost four times that of the unweighted estimate. To control for extreme weights, the authors truncated the weights by replacing outlying values with smaller ones. A new estimate of the effect of HIV treatment was obtained (25; 95% CI (11.6;38.9)). The authors further suggested to model the treatment assignment in a flexible way, e.g. including splines, to control for model misspecification.

Chapter 2

Background

2.1 Weighting in survey sampling

The general goal of weighting in survey sampling is to find a set of weights that corrects for disproportions in the sample, thus making the sample representative of a target population of interest. We may be interested in estimating a parameter, such as the mean, of this target population. To provide an example on how weighting is used, let us consider sampling two samples of sizes n_1 and n_2 from two strata, S_1 and S_2 of known sizes N_1 and N_2. Suppose we are interested in estimating the population total t_Y. A natural estimator for t_Y is the sample mean weighted by each population total, i.e. $\hat{t}_Y = N_1\bar{y}_1 + N_2\bar{y}_2$. This expression can also be written as,

$$\begin{aligned}
\hat{t}_Y &= N_1 \sum_{i \in S_1} \frac{y_i}{n_1} + N_2 \sum_{i \in S_2} \frac{y_i}{n_2} && (2.1)\\
&= \sum_{i \in S_1} \frac{y_i}{n_1/N_1} + \sum_{i \in S_2} \frac{y_i}{n_2/N_2} && (2.2)\\
&= \sum_{i} \frac{y_i}{\pi_i} && (2.3)
\end{aligned}$$

where $\pi_i = n_j/N_j$ if $i \in S_j$ and $j \in 1, 2$ are the inclusion probabilities. By weighting each sampled y_i for the inverse of its probability of selection, we obtain an unbiased estimate of the population total. The estimator $\hat{t}_Y$ is called the Horvitz-Thompson (HT) estimator (Horvitz and Thompson, 1952). In this thesis, we refer to Y as a random variable, to y as its realization, and to $\mathbb{E}[Y]$ as the expected value of Y.

2.1.1 Calibration Estimators

Auxiliary information about the target population, such as population means or totals, obtained from sources different than the survey, such as census, is widely used in survey sampling to improve survey estimates (Foreman and Brewer, 1971; Deville and Särndal, 1992; Valliant et al., 2013). For example, Deville and Särndal (1992) proposed a family of calibration estimators in which auxiliary information is incorporated. These estimators use calibrated weights, defined as the closest weights, given a distance measure, to the original sampling design weights that satisfy a set of constraints called calibration equations. Specifically, let us consider a sample S selected from a finite population U and consider the sampling design weights $d_i = 1/\pi_i \ \forall \ i \in S$, and the HT estimator, $\hat{t}_Y = \sum_{i \in S} y_i/\pi_i = \sum_{i \in S} d_i y_i$. Let us further assume that we have auxiliary information in the form of the population total of $\mathbf{X}$, $t_{\mathbf{X}} = \sum_{i \in U} \mathbf{x}_i$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ is the observed vector of p auxiliary variables. A calibration estimator is the estimator, $\hat{t}_Y^c = \sum_{i \in S} w_i y_i$, with weights w_i, $\forall \ i \in S$ which minimize a measure of distance, $L(w, d)$ while satisfying the following calibrating equation,

$$\sum_{i \in S} w_i \mathbf{x}_i = t_{\mathbf{X}}. \tag{2.4}$$

One choice of distance $L(w, d)$ is the least-squares distance function $\sum_{i \in S}(w_i - d_i)^2/d_i$. Minimization leads to the calibrated weights,

$$w_i = d_i(1 + \mathbf{x}_i^T \lambda), \tag{2.5}$$

where λ is the vector of Lagrange multipliers determined from (2.4) and equal to $\lambda = (\mathbf{X}^T \mathbf{D} \mathbf{X})^{-1}(t_{\mathbf{X}} - \hat{t}_{\mathbf{X}})$. The resulting estimator of t_Y is

$$\hat{\mathbf{t}}_y = \hat{t}_Y + (t_{\mathbf{X}} - \hat{t}_{\mathbf{X}})^T \hat{\mathbf{B}}, \tag{2.6}$$

where $\hat{\mathbf{B}} = (\mathbf{X}^T \mathbf{D} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{D} y$ is the solution to the weighted least square estimator, and $\hat{t}_{\mathbf{X}} = \mathbf{X}^T \mathbf{D}$. Calibration estimators provide a different derivation of the generalized regression estimator (Cassel et al., 1976; Särndal, 1980) in which auxiliary information is included. In addition, calibration estimators provide an example of the use of mathematical programming to obtain weights that satisfy specific constraints. Lumley et al. (2011) showed a connection between calibration estimators and the augmented inverse probability weighting estimator for missing data and causal inference proposed by Robins et al. (1994).

2.2 Missing data

Weighting can also be used to control for missing data. By missing data, we mean data values that are unobserved. Examples include non-response in survey sampling, when randomly chosen individuals provide partial or no answer, and dropout/non-compliance in randomized clinical trials, where individuals initially randomized to one treatment, do not show up for any clinical visits or occasionally miss visits. Missing data leads to samples that are non-representative of the target population, thus introducing selection bias. We now provide an example on how weighting is employed to control for this type of bias. Consider a randomized trial, in which units are randomly assigned to a treatment or placebo. Specifically, let $A_i = 1$ denotes the units assigned to the treatment, and $A_i = 0$

those assigned to placebo, for all $i = 1, \ldots, n$, where n is the sample size. Let Y_i be the outcome under study. The parameter of interest is the average treatment effect (ATE), $\Delta = \mu_1 - \mu_0$, where μ_t, $t \in \{0, 1\}$ is the mean outcome among treated and placebo individuals. Let us further define *full data*, the data we would have liked to have collected, *observed data*, the data actually observed, and *complete data*, the subset of the sample in which complete, non-missing, information is available. Having access to the full data, we would allow estimating Δ with the unbiased estimator $\hat{\Delta} = \hat{\mu}_1 - \hat{\mu}_0$. Let us now denote by R_i, $i = 1, \ldots, n$, the indicator of complete data, with $R_i = 1$ meaning that the i-th measurement was taken, and $R_i = 0$ meaning that it was missing. In this specific scenario, we would observe (A_i, R_i, R_iY_i), $i = 1, \ldots, n$. By focusing our attention on the treated patients, i.e. $A_i = 1$, the observed data are denoted by (R_i, R_iY_{1i}), $i = 1, \ldots, n_1$, where Y_{1i} is the i-th observed outcome among the treated. Under the assumption that the data are missing completely at random (MCAR), i.e. the probability of being missing (or observed) does not depend on Y_{1i}[1], the complete-case (CC) estimator $\hat{\mu}_{1CC} = \sum_{i=1}^{n_1} R_iY_{1i} / \sum_{i=1}^{n_1} R_i$ is an unbiased estimator for the ATE. When non-missing auxiliary information about p auxiliary variables, denoted by $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})$, $i = 1, \ldots, n$ is available, we can relax the MCAR assumption, with that of missing at random (MAR), defined as $R_i \perp\!\!\!\perp Y_{i1} | \mathbf{X}_i$. MAR means that the missingness mechanism is described in some way by the observed information $\mathbf{X}_i$. Under MAR, weighting by the inverse of the probability of being a complete case suggest the following inverse probability weighted complete case (IPWCC) estimator (Tsiatis, 2007; Robins et al., 1994),

$$\hat{\mu}_{1,IPW} = n_1^{-1} \sum_{n=1}^{n_1} \frac{r_i y_{1i}}{\pi(\mathbf{x}_i)} \tag{2.7}$$

where $\pi(\mathbf{X}_i) = P(R_i = 1|\mathbf{X}_i, Y_{i1}) = P(R_i = 1|\mathbf{X}_i)$[2] is the probability that the i-th randomly selected individual with auxiliary information $\mathbf{X}_i$, has complete data, and $\pi(\mathbf{x}_i)$ is its realization. The intuition behind the IPWCC estimator is that a sampled individual, with probability of having complete data equal to $\pi(\mathbf{X}_i)$, represents $1/\pi(\mathbf{X}_i)$ individuals from the population. Usually, $\pi(\mathbf{X}_i)$ is estimated from the data by using a logistic regression or a classification algorithm (Lee et al., 2010; Friedman et al., 2001), and it must be bounded away from zero for all values of $\mathbf{X}_i$. Similar assumptions and conditions are employed in the literature of causal inference under the name of ignorability and positivity, respectively (Imbens and Rubin, 2015). More on inverse probability weighting and missing data can be found in Seaman and White (2013).

2.3 Causal Inference

Most of the studies in science aim at answering causal rather than associational questions. For example, a medical doctor is interested in the effect of a treatment on some clinical out-

[1] MCAR: $R_i \perp\!\!\!\perp Y_{i1}$ and $P(R_i = 1|Y_{i1}) = \pi$, for all $i = 1, \ldots, n$.

[2] follows from MAR

comes, while a policy maker is interested in the effect of a new policy compared with an old one. While randomized experiments provide unbiased estimates of the ATE, economic and ethical limitations make them not viable with only observational data available. Although potentially huge in sample size, these data sets are observational, where causal effects are hidden by confounding factors which must be controlled for. The following are some known methods to control for confounding and estimate causal effects with observational data.

2.3.1 Inverse probability of treatment weighting

Potential outcomes provide a framework in which causal effects can be estimated from observational data (Rosenbaum and Rubin, 1983). A potential outcome $Y(a)$ is the outcome we would see if we were to be treated with treatment a. Specifically, let T denote an indicator variable defining treatment allocation ($A = 1$ if treated, and $A = 0$ if control), and let $\mathbf{X}$ be a vector of observed variables measured before treatment A was assigned. Let Y be the observed response variable formalized as

$$Y = Y(1)A + (1 - A)Y(0) \tag{2.8}$$

where the random variables $Y(1)$ and $Y(0)$ respectively represent the potential outcomes of a treated and of an untreated individual. Let $\Delta = E(Y(1) - Y(0)) = \mu_1 - \mu_0$ indicate the causal parameter of interest, the ATE, and μ_1 and μ_0 indicate the true means under each treatment. Under the assumption of

- *consistency*: the observed outcome corresponds to the potential outcome of applying treatment a, i.e. $Y = Y(a)$ if $A = a$,
- *positivity*[3]: the probability of getting treated given covariates is bounded away from zero and one, i.e. $\pi(\mathbf{X}) = P(A = 1|\mathbf{X}), 0 < \pi(\mathbf{X}) < 1$,
- *ignorability*[4]: the potential outcome is independent to the treatment assignment mechanism given covariates, i.e. $(Y(0), Y(1)) \perp\!\!\!\perp A|\mathbf{X}$,

inverse probability of treatment weighting provides unbiased estimates of the ATE when using observational data. In particular, it can be shown that

$$\hat{\Delta}_{IPW} = n^{-1} \sum_{i=1}^{n} \left(\frac{a_i y_i}{\pi(\mathbf{x}_i)} - \frac{(1 - a_i) y_i}{(1 - \pi(\mathbf{x}_i))} \right) = \mu_1 - \mu_0 \tag{2.9}$$

where $\pi(\mathbf{x}_i)$ is the observed realization of the propensity score for the i-th individual. Models, such as logistic regression, are used to estimate the unknown probability $\pi(\mathbf{x}_i)$. In case these models are incorrect, i.e. misspecification of the treatment assignment model, the IPW-estimator in (2.9) is biased. Augmented inverse probability weighting estimators

[3] also referred to as experimental treatment assignment or strict overlap

[4] also referred to as unconfundeness or exchangeability

(AIPW) (Scharfstein et al., 1999) have been proposed to overcome this issue. The key idea of AIPW is to combine regression and IPW estimators in one doubly robust (DR) estimator (Bang and Robins, 2005). DR refers to the fact that unbiased estimate of the ATE can be obtained whenever either of the outcome model or the propensity score model is correctly specified. Authors showed, however, that DR estimators may yield highly biased inferences when neither of the two models is correctly specified (Kang and Schafer, 2007a). Inverse probability weighting is also used to consistently estimate the parameters of a marginal structural model (MSM) (Robins, 2000). MSM are causal models in the form, for example, of $E(Y(a)) = \beta_0 + \beta_1 a$, where β_1 is the causal parameter of interest. Weighted ordinary least square estimator is tipically used to estimate β_1. MSM have been extensively used to estimate the causal effect of a time-varying treatment on an outcome of interest, such as, for example, time to death (Robins et al., 1999; Robins, 2000; Hernán et al., 2001).

2.3.2 Matching

Matching is a widely used technique to control for confounding in observational studies (Stuart et al., 2011b). The essential idea of matching is to assign one or more controls with similar observed covariates to each treated unit, thus balancing their distributions and, consequently, control for confounding. In the past decades, several matching methods have been developed, and the literature on the topic is vast (see Stuart et al. (2011b) for a review). Several measures of distance are used to evaluate if a control is a good match for a treated unit. Commonly used measures are

- exact, $D_{ij} = \{0, \text{ if } X_i = X_j; \infty, \text{ if } X_i = X_j\}$;
- Mahalanobis, $D_{ij} = (X_i - X_j)^T \Sigma^{-1} (X_i - X_j)$, where Σ is the variance covariance matrix in the full control group when estimating the average treatment effect on the treated (ATT) and that of X in the full control and pooled treatment groups when estimating the ATE; and
- Propensity score, $D_{ij} = |\pi(X_i) - \pi(X_j)|$ where $\pi(X_i)$ is the propensity score for the i-th individual.

"Greedy" and optimal matching methods are the most popular used matching algorithms. An example of a greedy matching method is the nearest neighbor matching method, in which a control individual, with the smallest distance from the treated, is selected for each treated individual. Rosenbaum (2002), showed how greedy methods can perform poorly as matching methods. In contrast, optimal matching (Rosenbaum, 1989), proposes to choose individual matches by minimizing a global measure of distance. This method is related to the literature of network flow theory, in which the standard problem is to find a flow of minimal cost in a network. In particular, Rosenbaum (1989) showed that optimal matching can be reinterpreted as a "personnel assignment" mathematical programming problem (Kuhn, 1955). Recently, Kallus (2016) proposed a class of generalized optimal matching methods, which includes many existing matching methods such as nearest-neighbor matching, optimal caliper matching, 1:1 matching, coarsened exact matching, and various near-fine balance approaches.

2.3.3 Covariate balancing

The ultimate goal of IPW and matching is to balance covariates in order to obtain an unbiased estimate of the causal parameter of interest. Methods that combine the spirit of probability weighting and matching have received recent attention. Kallus (2016) reinterpreted optimal matching by providing balancing weights that balance covariates. Imai and Ratkovic (2014), proposed covariate balancing propensity score in which covariate balance is optimized. Hainmueller (2012) presented entropy balancing method, in which each observation is weighted to achieve optimal balance and the Kullback-Leibler divergence from a set of target weights is minimized. Chan et al. (2016) presented a general class of calibration estimators, which are constructed to attain exact three-way balance. Zubizarreta (2015) proposed a set of weights that are stabilized, i.e. have minimum variance while balancing covariates. Li et al. (2017) proposed a set of new weighting schemes which balance covariates via propensity ccore weighting. Athey et al. (2016) developed a method that allows sparse regression methods to be used to estimate ATE in high-dimensional linear models. Hirshberg and Wager (2017) proposed a method that efficiently estimates treatment effects by using weights that directly optimize worst-case risk bounds. Wong and Chan (2017) proposed a method that uniformly approximate covariate balance for functions in a reproducing-kernel Hilbert space. Zhao (2016) proposed to balance covariates by using tailored loss functions.

2.3.4 Causal inference in longitudinal studies

In longitudinal studies, where T measures are collected for each unit $i = 1, \dots, n$, the causal effect of a time-varying treatment on an outcome of interest can be estimated. Standard methods such as regression adjustment or matching fail to provide consistent estimate of the causal effect in presence of time-dependent confounders (Robins, 2000; Blackwell, 2013). Time-dependent confounders are time-varying factors that are affected by previous treatments and affect future ones (Robins, 2000). A common example of the role of time-dependent confounder is given by CD4 cell count in the study of the effect of HIV treatment on mortality among HIV-infected patients (Hernán et al., 2000, 2001; HIV-Causal Collaboration et al., 2010, 2011; Lodi et al., 2017). CD4 cell count is both an independent predictor of initiation of HIV treatment and survival as well as being itself influenced by prior HIV treatment. Methods to deal with time-dependent confounding have been proposed in the statistical literature (Daniel et al., 2013). Among others, MSM have been used to estimate the causal effect of a time-dependent treatment on an outcome of interest. For each $\overline{a}_T$, we define the MSM for the effect of a time-varying treatment on the mean of Y as follows,

$$\mathbb{E}\left[Y(\overline{a}_T)\right] = g(\overline{a}_T, \beta) \tag{2.10}$$

where $g(\overline{a}_T, \beta)$ is some known function, for example $g(\overline{a}_T, \beta) = \beta_1 + \Delta \sum_{t=1}^{T}(a_t)$, the parameter Δ is the causal parameter of interest and $Y(\overline{a}_T)$ is the potential outcome if the unit were to be treated with treatment regime $\overline{a}_T$. Under consistency, positivity and sequential ignorability (Imbens and Rubin, 2015; Hernan and Robins, 2010), IPW is used to consistently estimate the parameters of the MSM. Inverse probability of treatment weights are defined as

$$w(\overline{a}_T, \overline{x}_T) = \prod_{t=1}^{T} \frac{h(\overline{A}_t)}{P(A_t = a_t|\overline{A}_{t-1} = \overline{a}_{t-1}, \overline{X}_t = \overline{x}_t)} \tag{2.11}$$

where A_t is the binary treatment variable at time t, X_t the time-dependent confounders at time t, $\overline{A}_t$ is the treatment history up to time t, $\overline{X}_t$ is the history of time-dependent confounders up to time t, and $h(\overline{A}_t)$ is a known function of the treatment history bounded between zero and one. The set of inverse probability weights is obtained by setting $h(\overline{A}_t) = 1$, while the set of stable inverse probability weights is obtained by setting $h(\overline{A}_t) = P(A_t = a_t|\overline{A}_{t-1} = \overline{a}_{t-1})$. Weights in the form of (2.11) are estimated by using parametric methods such as logistic regression, along with other methods based on the literature of statistical learning (Karim et al., 2017; Gruber et al., 2015; Karim and Platt, 2017). Other methods such as g-computation formula (Robins, 1986), and g-estimation of structural nested models (Robins, 2000) provide alternative solutions to the problem of time-dependent confounders. A review can be found in Daniel et al. (2013).

Matching for longitudinal data. Matching has also been used in the context of longitudinal data. Li et al. (2001) proposed optimal balanced risk set matching, in which a patient with similar history of symptoms up to time t is matched to a patient that receive treatment at time t. The method is based on the solution of a integer programming problem. Lu (2005) proposed a time-dependent propensity score used in risk set matching, which is based on the Cox proportional hazards model. It is not clear, however, if these methods can control for time-dependent confounding.

Covariate balancing for longitudinal data. Differently from inverse probability weights, Imai and Ratkovic (2015) proposed to estimate weights by generalizing the covariate balancing propensity score (CBPS) methodology. CBPS estimates robust inverse probability weights for MSM by optimizing covariate balance. However, CBPS does not control for informative censoring.

2.4 Limitations of probability weighting methods

As described in the previous sections, probability weighting methods are widely used in statistics. It is well known, however, that they have limitations. These limitations are: (1) statistical inference may be inefficient when weights contain extreme values; (2) statistical inference may be highly biased in case of model misspecifications.

2.4.1 Extreme probability weights

In today's medical and epidemiological research, the most popular approach to deal with extreme weights is truncation, which consists of replacing outlying weights with smaller weights. For example, all weights above the 99th percentile of their sample distribution may be replaced with the 99th percentile itself. Methods have been proposed to obtain optimal cutoff points (Potter, 1990; Cox and McGrath, 1981; Kokic and Bell, 1994; Rivest et al., 1995; Hulliger, 1995). Approaches other than truncation, have been proposed (Pfeffermann and Sverchkov, 1999; Beaumont, 2008; Beaumont et al., 2013; Elliot and Little, 2000; Elliott, 2009; Zubizarreta, 2015). In longitudinal studies, truncation remains the only method used to control for extreme weights (Cole and Hernán, 2008; Xiao et al., 2013).

2.4.2 Model misspecification

The propensity score needs to be estimated from the data. Several authors showed that in case of misspecification of the missing/treatment mechanism model, biased estimates of the parameter of interest are obtained. The problem is exacerbate in longitudinal study, where probability weights are multiplied across time points. Covariate balance methods, introduced in Section 2.3.3, mitigate the effect of possible model misspecification of the parametric model for the propensity score by obtaining weights that maximize the resulting covariate balance. To fix ideas, under consistency, positivity and ignorability and by the law of total probability (LTE), consider $\Delta = E(Y(1) - Y(0))$ the parameter of interest and consider the following decomposition based on the weighted average,

$$\begin{aligned}
\mathbb{E}\left[W\mathbb{1}[A=a]Y\right] &= \mathbb{E}\left[W\mathbb{1}[A=a]Y(a)\right] \quad \text{(by consistency)} \\
&= \mathbb{E}\left[W\mathbb{E}\left[\mathbb{1}[A=a]Y(a)|\mathbf{X}\right]\right] \quad \text{(by LTE)} \\
&= \mathbb{E}\left[W\mathbb{E}\left[\mathbb{1}[A=a]|\mathbf{X}\right]\mathbb{E}\left[Y(a)|X\right]\right] \quad (\text{by } Y(a) \perp\!\!\!\perp T|\mathbf{X}) \\
&= \mathbb{E}\left[\mathbb{E}\left[Y(a)|\mathbf{X}\right]\right] + \delta_a \\
&= \mathbb{E}\left[Y(a)\right] + \delta_a
\end{aligned} \tag{2.12}$$

where $\mathbb{E}\left[\mathbb{1}[A=a]|X\right]$ is the propensity score. One way to obtain an unbiased estimate of Δ is to use inverse probability weights $W = 1/\mathbb{E}\left[\mathbb{1}[A=a]|\mathbf{X}\right]$. A more robust[5] way is to

[5] that does not require a specification of the treatment assignment model

find a set of weights that makes $\delta_a = 0,\ a \in 0,1$ by balancing $\mathbb{E}\left[Y(a)|\mathbf{X}\right]$ across treated and controls.

2.5 Kernels and Gaussian processes

In this section we briefly introduce the concepts of kernels, reproducing kernel Hilbert spaces and Gaussian processes for machine learning.

2.5.1 Kernels and reproducing kernel Hilbert spaces

A kernel is a function that quantifies the similarities between observations and refers to a dot product between observed characteristics of the individual, referred to as features. Specifically, a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is defined as a kernel if there exist a map $\phi : \mathcal{X} \to \mathcal{H}$ and an $\mathbb{R}$-Hilbert space such that $\forall\ x, x^\top \in \mathcal{X}$,

$$k(x, x^\top) := \left\langle \phi(x), \phi(x^\top) \right\rangle_{\mathcal{H}}. \tag{2.13}$$

For a formal definition of Hilbert space and inner product see Schölkopf and Smola (2002). All kernels are positive definite functions. Typical kernels are,

- *Linear kernel:* $k(x, x^\top) = x \cdot x^\top$.
- *Polynomial kernel:* $k(x, x^\top) = (x \cdot x^\top + 1)^d, \quad d \in \mathbf{N}$.
- *Gaussian kernel:* $k(x, x^\top) = e^{-\frac{\|x - x^\top\|^2}{\sigma^2}}, \quad \sigma > 0$.

The space of functions defined by a kernel on the feature space $\mathcal{X}$ is known as reproducing kernel Hilbert spaces (RKHS). RKHS satisfy, among others, the reproducing property, $\forall\ x \in \mathcal{X}, \forall\ f \in \mathcal{H},\ \langle f, k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$, i.e. the evaluation of f at x can be interpreted as an inner product in feature spaces. RKHS are used in supervised learning, such as classification, in which the task is to choose where a new observation belongs between two or more categories, e.g. support vector machines (Schölkopf and Smola, 2002). The key idea is to transform the data from the observed features space $\mathcal{X}$ into a higher dimensional space $\mathcal{H}$, thus allowing to linearly separate the data. This idea refers to as the "kernel trick" (Hofmann et al., 2008).

2.5.2 Gaussian processes for machine learning

A Gaussian process (GP) is a stochastic process in which any finite finite number of random variables has a joint normal distribution (Rasmussen, 2004). A GP is completely specified by its mean and covariance function, defined as

$$\begin{aligned} m(\mathbf{x}) &= \mathbb{E}[f(\mathbf{x})] \\ k(\mathbf{x}, \mathbf{x}^\top) &= \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))\left(f(\mathbf{x}^\top) - m(\mathbf{x}^\top)\right)] \end{aligned} \tag{2.14}$$

and it is usually defined as $f(\mathbf{x}) \sim \mathcal{GP}\left(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}^\top)\right)$. A GP can be viewed as a machine learning prediction algorithm. For example, the Bayesian linear regression model $f(\mathbf{x}) = \phi(\mathbf{x})^T\beta$ with prior $\beta \sim \mathcal{N}(\mathbf{0}, \Sigma_p), \Sigma_p = \mathbb{E}[\beta\beta^T]$ with mean and covariance

$$\begin{aligned} \mathbb{E}[f(\mathbf{x})] &= \phi(\mathbf{x})^T\mathbb{E}[\beta] = 0 \\ \mathbb{E}[f(\mathbf{x})f(\mathbf{x}')] &= \phi(\mathbf{x})^T\Sigma_p\phi(\mathbf{x}') \end{aligned} \tag{2.15}$$

is a GP. Based on the results presented in Section 2.5.1, the covariance function can be obtained by using kernels. Different choices of the kernel are employed for different prediction problems, such as regression and classification. Once a kernel has been chosen, its hyperparameters, the parameters for which the covariance function depends on, are usually obtained by minimizing the log-likelihood with respect to those parameters. More on GP for machine learning can be found in Rasmussen (2004).

2.6 Mathematical Programming

In general terms, mathematical programming, also refereed to as optimization, is the task of minimizing a risk function with or without constraints (Griva et al., 2009). This topic covers minimization of one variable, convex minimization, maximization problems, constrained optimization, and non-convex optimization. Let p be a vector of parameters in $\mathbb{R}^k$, a general form of constrained optimization problem follows,

$$\begin{aligned} \underset{p \in \mathbb{R}^k}{\text{minimize}} \quad & f(p) \\ \text{subject to} \quad & g_i(p) \le 0, \quad i \in \mathcal{I}, \quad \mathcal{I} \cup \varepsilon = \{1, \dots, m\} \\ & h_i(p) = 0, \quad i \in \varepsilon, \quad \mathcal{I} \cap \varepsilon = \emptyset \\ & p \ge 0 \end{aligned} \tag{2.16}$$

where $f : \mathbb{R}^k \mapsto \mathbb{R}$, $g : \mathbb{R}^k \mapsto \mathbb{R}^m$ and $h : \mathbb{R}^k \mapsto \mathbb{R}^n$, and where $p \ge 0$ are k bound constraints. In this thesis we will mainly consider problems in the form of (2.16) with $f(p)$ stricly convex and, when present, g_i convex, thus admitting unique solution. Given some regularity conditions, the Karush Kuhn Tucker (KKT) conditions are first-order necessary condition for a solution of problems in the form of (2.16). KKT generalizes the method of Lagrange multipliers to linear constraints. Apart of few special cases, in which analitical solutions can be found, the system of inequalities and equations obtained from the KKT conditions is solved by using optimization algorithms. Examples include, quadratic programming methods, which are used in Paper II and III and the primal-dual interior point method, which is used in Paper I. Algorithms are implemented in readily available software, such as the R packages Gurobi and IPOPTR.

Chapter 3

Aims of the thesis

The overall aim of this thesis was to develop, investigate, and apply novel methods to estimate optimal probability weights.

More specifically, it aimed

- to develop a method to estimate optimal probability weights that controls for the precision of the resulting weighted estimator;
- to extend optimal probability weights when interested in the estimation of the causal effect of a time-varying treatment on a survival outcome;
- to develop a novel method that simultaneously balances time-dependent confounders and control for the precision of the resulting marginal structural model estimate in longitudinal studies;
- to show the applicability of the proposed methods by using data from the HIV literature.

Chapter 4

Summary of the studies

4.1 Paper I

The aim of Paper I was to develop a method to estimate optimal probability weights. Optimal probability weights are the solution to a constrained optimization problem that minimizes the Euclidean distance from the target (original/untruncated) weights among all sets of weights that satisfy a constraint on the precision of the resulting weighted estimator. Specifically, let $\hat{\theta}_{w^*}$ be an unbiased estimator for a population parameter θ^* that uses weights $w^* = (w_1^*, \ldots, w_n^*)^T$, with $\mathbf{1}^T w^* = 1$ and $w^* \geq 0$. The set of w^* can be, for example, the set of inverse probability weights used to control for missing data or confounding. When w^* contains outliers, the standard error σ_{w^*} may be large and inference on θ^* inefficient. We suggest deriving the weights $\hat{w}$ that are closest to w^* with respect to the Euclidean norm $\|w - w^*\|$, under the constraint that the estimated standard error $\hat{\sigma}_{\hat{w}}$ be less than or equal to a specified constant $\xi > 0$. The corresponding constrained optimization problem can be written as follows,

$$\begin{aligned}
\underset{w \in \mathbb{R}^n}{\text{minimize}} \quad & \|w - w^*\|_2 && (4.1)\\
\text{subject to} \quad & \hat{\sigma}_w \leq \xi && (4.2)\\
& w \leq \epsilon && (4.3)\\
& w \geq 0. && (4.4)
\end{aligned}$$

Constraint (4.9) guarantees that the estimated standard error of the estimator with weights $\hat{w}$ is less than or equal to ξ. Constraints (4.10) and (4.11) guarantee that the optimal weights $\hat{w}$ are bounded and non-negative, respectively. In Section 3 of Paper I, we described how the Lagrange multipliers and objective function from the optimization problem can help assess the trade-off between bias and precision of the weighted estimator. In a simulation study, we showed that optimal probability weights performed better than truncated weights with respect to bias and mean squared error of weighted least-squares regression coefficients. We also showed that optimal probability weights often led to large gains in precision at the cost of small bias. We illustrated the use of optimal probability weights in an analysis of the effect of timing of treatment initiation on long-term health outcome in patients infected by HIV, using data from the Swedish InfCare HIV registry. Our findings indicated that the age at the start of treatment was a relevant effect modifier, and correct timing of treatment initiation was more important in younger patients.

4.1.1 Additional results

In addition to the results provided in Paper I, we also constructed 95% confidence intervals for all the scenarios reported in Figures 1 and 2 of Paper I. The observed coverage is shown in Figures 4.1-4.4 below. In all scenarios, when the optimal weights are close to the target weights, the bias of the optimal estimator, $\hat{\theta}_{\hat{w}}$, with respect to the target parameter, θ_{w^*}, is slight and the observed coverage of the confidence interval is near the nominal level. As the optimal weights move away from the target weights, the bias increases and the coverage decreases. The coverage of the truncated estimator, $\hat{\theta}_{\overline{w}}$, is always worse than that of the optimal estimator, $\hat{\theta}_{\hat{w}}$, especially in the presence of a strong confounding effect (Figure 4.3).

4.2 Paper II

The aim of Paper II was to extend optimal probability weights to estimating the causal effect of a time-varying treatment on a survival outcome. As shown by Hernán et al. (2001), inverse probability of treatment and censoring weighting is used to consistently estimate the parameters of the marginal structural Cox model. However, these methods rely substantially on the positivity assumption. Practical violations of the positivity assumption are common with survival data, resulting in extreme weights, low precision and erroneous inferences. As presented in Paper I, truncation is the most used technique to control for extreme weights. In Paper II, we proposed to obtain weights $\hat{w}_o^{(t)}$ that are the closest to $\hat{w}_*^{(t)}$, the estimated set of target (original/untruncated) weights, with respect to the Euclidean norm, while constraining the variance of the weights $\hat{w}_o^{(t)}$ to be less or equal to a specified level ξ. The resulting quadratic optimization problem can be formulated as follows,

$$\begin{aligned} \underset{w_o^{(t)} \in \mathbb{R}^{n \times t}}{\text{minimize}} \quad & \|w_o^{(t)} - \hat{w}_*^{(t)}\|_2 && (4.5) \\ \text{subject to} \quad & \|w_o^{(t)} - \overline{w}_o^{(t)}\|_2^2 \leq \xi && (4.6) \\ & w_o^{(t)} \geq 0 && (4.7) \end{aligned}$$

where $\overline{w}_o^{(t)}$ is the mean of the weights $w_o^{(t)}$. Differently from Paper I, where we constrained the standard error of the resulting weighted estimator, in Paper II we constrained the variance of the weights. This formulation of the optimization problem is novel and has two main advantages: (1) it is quadratic and convex and therefore admits a unique solution; and (2) it is independent of both the chosen estimator for the causal parameter of interest

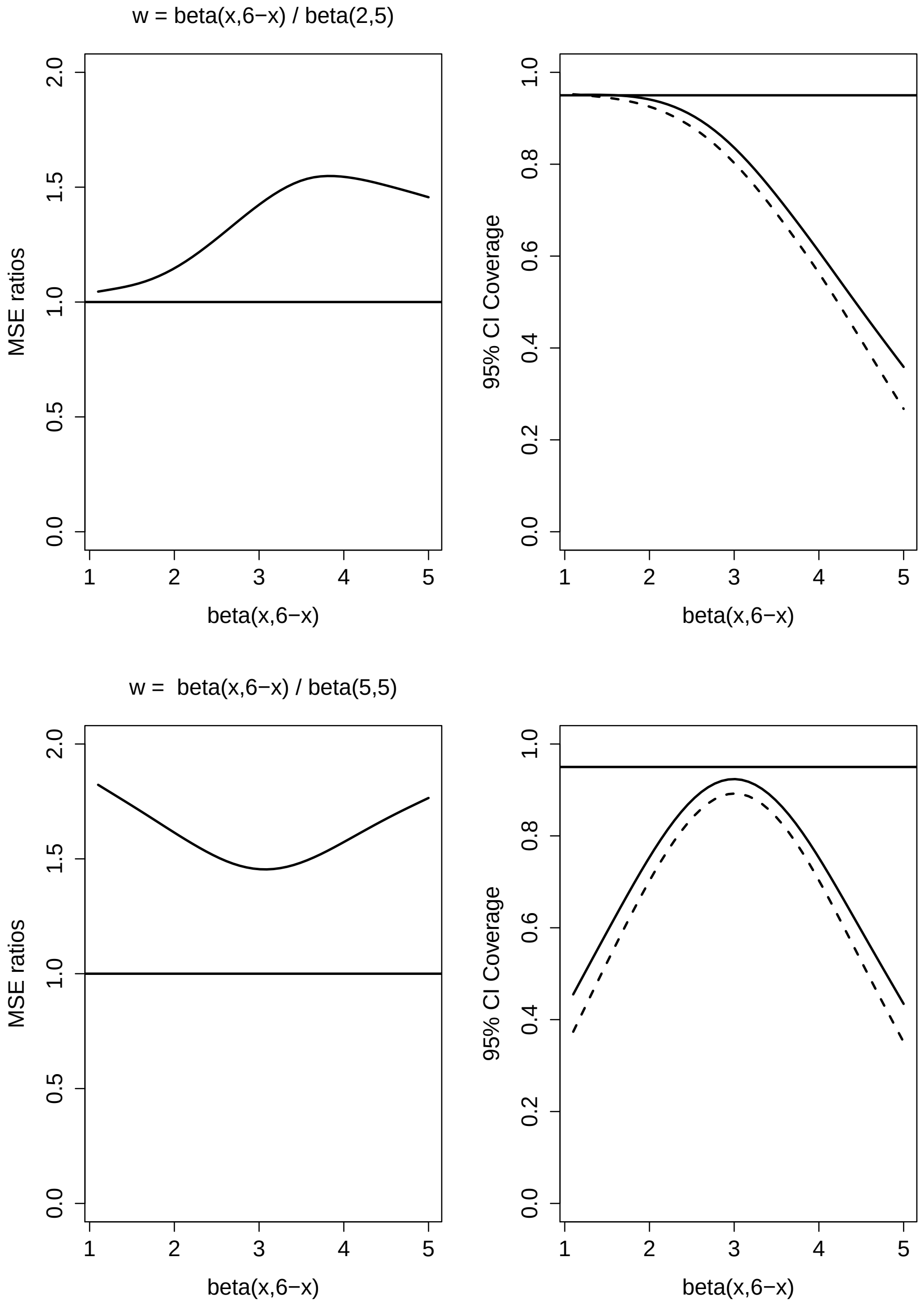

Figure 4.1: The right-hand-side panels show the observed coverage of 95% confidence intervals using optimal weights (solid lines) and truncated weights (dotted) corresponding to the scenarios considered in the left-hand-side panels of Figure 1 of the manuscript, which are reported in the left-hand-side panels above.

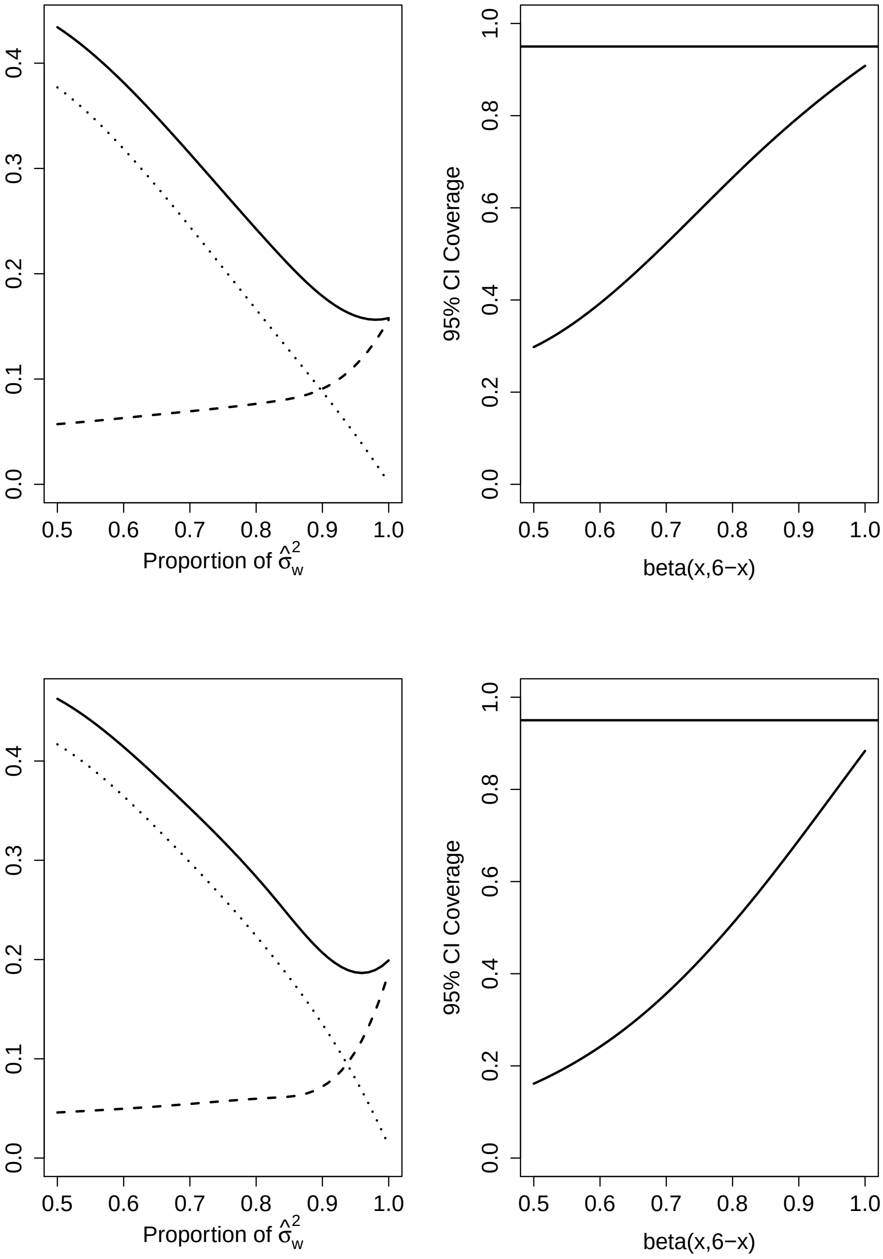

Figure 4.2: Observed coverage of 95% confidence intervals (right-hand-side panels) corresponding to the scenarios considered in the right-hand-side panels of Figure 1 of the manuscript, which are reported in the left-hand-side panels above.

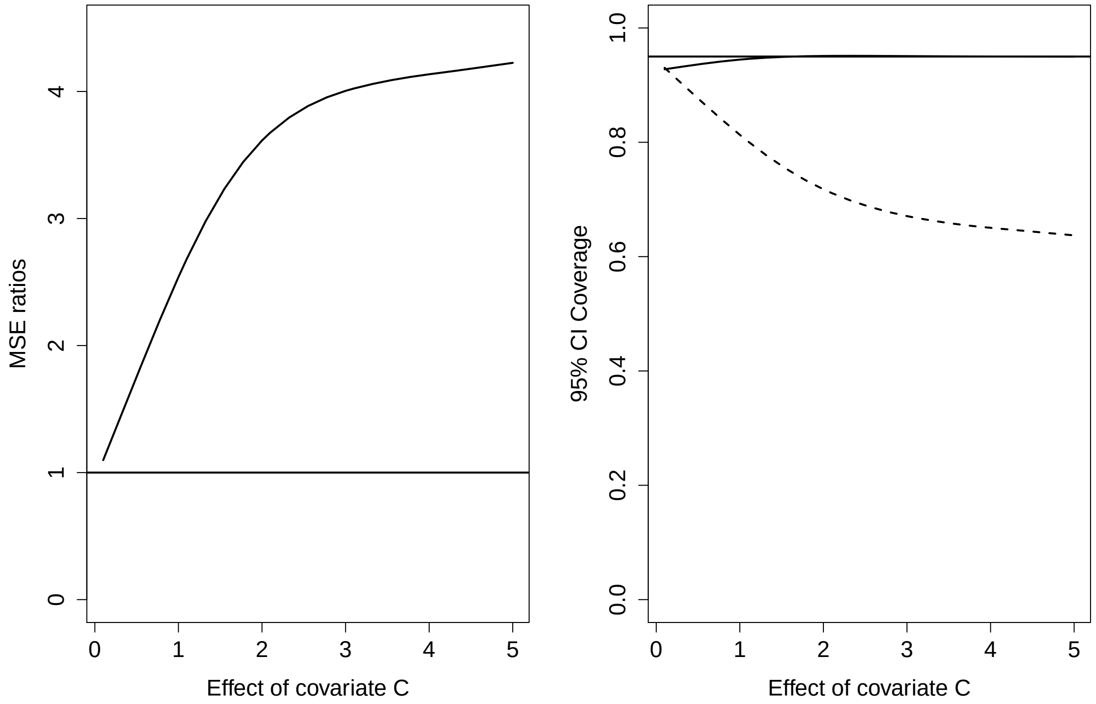

Figure 4.3: The right-hand-side panel shows the observed coverage of 95% confidence intervals using optimal weights (solid lines) and truncated weights (dotted) corresponding to the scenarios considered in the left-hand-side panel of Figure 2 of the manuscript, which is reported in the left-hand-side panel above.

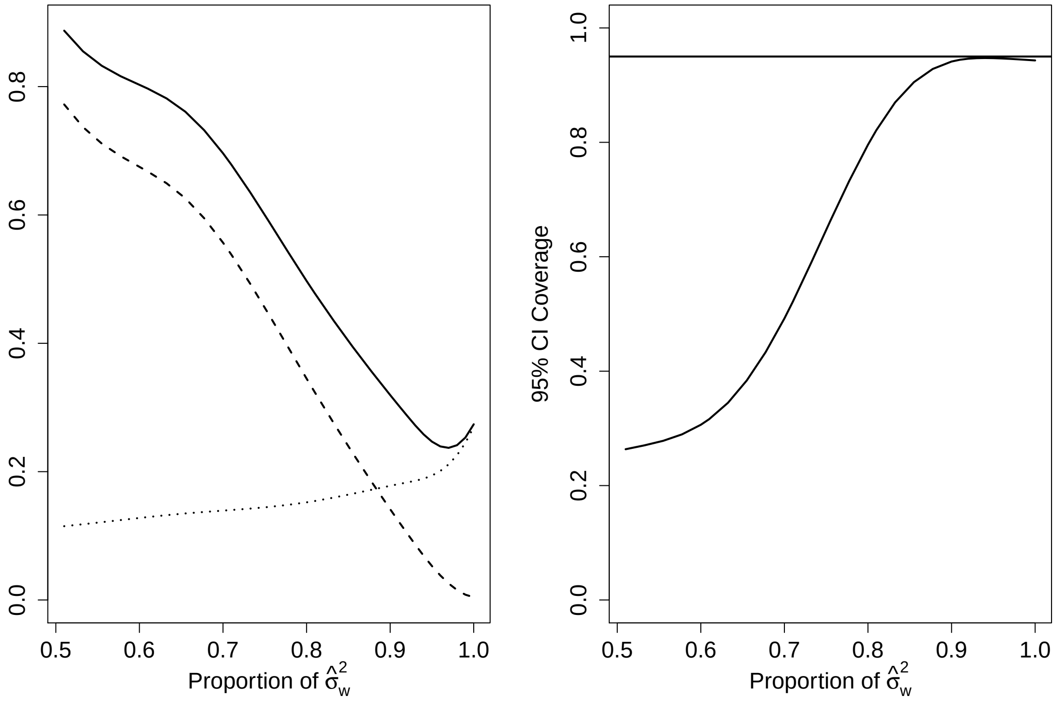

Figure 4.4: Observed coverage of 95% confidence intervals (right-hand-side panel) corresponding to the scenarios considered in the right-hand-side panel of Figure 2 of the manuscript, which are reported in the left-hand-side panel above.

and that for its standard error. The optimization problem in (4.8), can be reformulated as a unconstrained problem in which we shrink/penalize the variance of the weights by a penalization parameter λ. In a simulation study we showed how the bias and mean square error of the causal effect of a time-varying treatment that used the proposed weights outperformed the truncated weights across all the considered truncation levels, especially at high percentiles. We illustrated the use of the proposed weights to evaluate the causal effect of treatment initiation on time to death among people who live with HIV, using data from the Swedish InfCare HIV registry. We concluded that people who live with HIV can benefit from being treated.

4.3 Paper III

The aim of Paper III was to develop a novel method that provides weights that optimally balance time-dependent confounders while controlling for the precision of the resulting MSM estimate. Specifically, MSM have been widely used to estimate the causal effect of a time-varying treatment on an outcome in the presence of time-dependent confounding. These methods, however, have two main limitations: (1) as shown in Paper I and II, they heavily rely on the positivity assumption, and (2) they are highly sensitive to the misspecifiation of the treatment assignment model. Various statistical methods have been proposed in an attempt to overcome these challenges. To control for misspecification of the treatment assignment model, Imai and Ratkovic (2015) proposed covariate balance propensity score (CBPS), which estimates robust inverse probability weights for MSM by optimizing covariate balance. The method ensures that the first moment of each covariate is balanced even in case of model misspecification (Imai and Ratkovic, 2014). As shown in Paper I and II, to control for practical violation of the positivity assumption, several authors (Cole and Hernán, 2008; Xiao et al., 2013) suggested truncation, which consists of replacing outlying weights with less extreme one. In Paper II, we proposed to find weights that minimize the Euclidean distance from the estimated inverse probability weights while constraining the variance of the weights. To our knowledge, no methods have been proposed which simultaneously optimize covariate balance and control for the precision of the resulting MSM estimate. In Paper III, we proposed Kernel Optimal Weighting (KOW), which provides weights that optimally balance time-dependent confounders while controlling for the precision of the resulting MSM estimate by directly minimizing the error in estimation. This error is expressed as an operator derived from the g-computation formula (Robins, 1986) and KOW minimizes its operator norm with respect to a reproducing kernel Hilbert spaces (RKHS) by solving a quadratic optimization problem. In Section 5 of Paper III, we showed the results of a simulation study aimed at comparing bias and mean square error (MSE) of the estimated cumulative effect of a time-varying treatment on an outcome of interest estimated by using KOW, inverse probability weighting (IPW), stable inverse probability weighting (SIPW) and covariate balancing propensity score (CBPS). In summary, the proposed KOW outperformed IPW, SIPW and CBPS with respect to

MSE across all sample sizes and simulation scenarios. We also presented two empirical applications of KOW. In the first, we estimated the effect of treatment initiation on time to death among people who live with HIV. In the second, we evaluated the impact of negative advertisement on vote shares.

4.4 Paper IV

In Paper IV, we evaluated the effect of therapy switch on time to second-line HIV treatment failure among people who live with HIV, using data from the Swedish InfCare HIV registry. We defined treatment failure as one viral load ≥ 200 copies/mL after at least six months of a new treatment line initiation. Switch to second-line treatment was defined as: (1) switch without treatment failure; (2) switch due to treatment failure; (3) switch due to treatment failure and detectable drug resistance mutation. In Paper IV, we found that there was a significant difference in time to second-line treatment failure between patients who switched without treatment failure and those who did. Paper IV has few methodological limitations. First, no methods, such as IPW or covariate balancing, aside of modelling the outcome model, was used to control for confounding. Second, although Laplace regression (Bottai and Zhang, 2010) represents a practical alternative for the estimation of conditional quantiles with censored data, the method fails to yield a consistent estimator when the Laplacian assumption does not hold (Koenker, 2011). In the following section, we introduce the results of an alternative analysis in which we used a weighted Cox regression model, weighted by optimal probability weights as presented in Paper I.

4.4.1 Additional results

In this section we present the results of an alternative analysis where we used optimal probability weights, as described in Paper I, to estimate the effect of switch to second-line treatment on time to second line treatment failure. We constructed the set of inverse probability weights considering the follow confounders: CD4 cell count (<200; 200–350; 350–500; and >500 cells/mL) and viral load ($\leq$ 100.000; >100.000 copies/mL) at baseline and at second-line HIV treatment initiation; type of treatment regimen (NNRTI based, PI/r based, PI based, and Other) at first and second-line HIV treatment; age (0–30; 31–40; 41–50; >50 years) at first-line HIV treatment initiation; route of transmission (heterosexual, men having sex with men, people who inject drugs, other); country of birth (Sweden vs Non-Sweden); and the interactions between age at baseline and CD4 cell count at baseline and at second-line HIV treatment initiation. We obtained the set of optimal probability weights (OPW) by solving the following constrained optimization problem,

$$
\begin{aligned}
\underset{w \in \mathbb{R}^n}{\text{minimize}} \quad & \|w - w^*\|_2 && (4.8) \\
\text{subject to} \quad & \hat{\sigma}_w \leq \xi && (4.9) \\
& w \leq \epsilon && (4.10) \\
& w \geq 0 && (4.11)
\end{aligned}
$$

where w^* is the set of IPW weights, and $\hat{\sigma}_w$ is the robust standard error (Freedman, 2006). We set ξ equal to the value corresponding to the 20% increase in precision with respect to the standard error obtained by using IPW. We used a marginal structural Cox models weighted by IPW and OPW, to estimate the hazard ratio of treatment switch on time to second-line treatment failure. Table 4.1 shows the results of our analysis. We used robust standard errors (Freedman, 2006).

Table 4.1: Estimated hazard ratio of treatment switch on time to second-line treatment failure by using OPW and IPW.

		OPW		IPTW	
	$\hat{HR}$	95% CI	λ	$\hat{HR}$	95% CI
Switch due to TF	2.26	(1.25;4.01)	8.18	2.83	(1.35;5.92)
Switch due to TF+DRM	2.66	(1.63;4.32)	82.71	2.65	(1.44;4.86)

Note: $\hat{HR}$ is the estimated hazard ration, CI is the confidence interval, λ is the Lagrange multiplier associated with constraint 4.9. TF means treatment failure and DRM means drug resistance mutation.

We conclude that the there is a statistically significant effect of treatment switch on time to second-line HIV treatment. OPW provided similar results as those obtained with IPW but with higher precision.

Chapter 5

Discussion

Methods based on probability weighting have been subject of intense research in the field of statistics in the past decades, with applications in epidemiology, economics, and medicine. However, as shown the previous sections, these methods have several limitations. In this Ph.D. thesis we developed, evaluated and applied novel methods based on mathematical programming techniques which aimed at providing optimal probability weights. The optimal probability weights proposed can be used to control for extreme weights in cross-sectional studies (Paper I), and in longitudinal studies (Paper II). In addition, we provided a novel method that simultaneously balance time-dependent confounders and control for extreme weights in longitudinal studies (Paper III). The method introduced in Paper III, emerged by the cross-pollination between causal inference, machine learning, and mathematical programming. This Ph.D. thesis work provided methods that will likely help to (1) extend the use of probability weighting methods in medicine, epidemiology, economics and social sciences, (2) extend knowledge on how mathematical programming and machine learning could be used to conduct robust analyses for improved decision-making, and (3) provide powerful, strong, and robust results to clinicians and policy-makers.

Chapter 6

Future research

Designing and developing novel methods for estimating causal effects which connects machine learning, statistics, personalization, and medicine has recently been recognized as an important and emerging field, with research groups worldwide studying and extending past knowledge regarding these methods. The following are ideas for future work.

- *Robust estimation of dynamic treatment regimes.* Dynamic treatment regimes (DTR)s generalize personalized treatments in settings where treatments are time-varying, tailoring decisions based on the time-varying state of individual patients. Future research may focus on the development of methods that robustly estimate optimal DTRs in presence of time-dependent confounding. For instance, the performance of the optimal probability weights proposed in this Ph.D. thesis could be evaluated with popular methods to compare and find optimal DTRs, such as marginal structural models, and with novel learning methods that build upon the literature of statistical learning, such as backward outcome weighted learning (Zhao et al., 2015). In addition, future research may focus on the possible extensions to the evaluation of quantile optimal DTRs. Finally, future work may investigate the challenges and potential solutions to the issue of conducting inference for optimal DTRs (Chakraborty and Murphy, 2014; Chakraborty et al., 2010).

- *Stable weights for longitudinal censored data.* Several methods have been developed to alleviate the presence of extreme weights when considering one single time point (Santacatterina et al., 2017; Zubizarreta, 2015, among others). However, when estimating the causal effect of a time-varying treatment with longitudinal observational data few alternatives are available. Additionally, in longitudinal observational data, the outcome of interest may happen to be unobserved, due for example to loss of follow up. When the reasons for these losses are related to the study, selection bias is introduced. This phenomenon is usually called informative censoring. Although already partially addressed in Paper III, future research may focus on providing optimal weights with longitudinal data affected by time-dependent confounding and informative censoring.

- *Optimally balanced quantile treatment effects.* The literature of causal inference has been focusing mainly on the average treatment effects. However, there is a vast area of problems where the interest is in investigating quantiles. In particular, formulating novel methods for DTRs in terms of quantiles allows the analyst to give a more comprehensive picture of the treatment effect on the outcome. Future research may focus on the development of methods that provide optimal weights to estimate quantile treatment effects.

- *Optimal covariate balance for non-binary treatments.* Medical therapies and interventions commonly involve multiple and continuous treatments. For instance, more than twenty-five antiretroviral drugs in six classes are available for the treatment of HIV infection. An example of continuous treatment is the evaluation of the effect of red and processed meat intake on the risk of cardiovascular disease or cancer (McAfee et al., 2010). Future research may focus on the development of optimal weights, such as those presented in Paper III, that account for non-binary and continous treatments. For instance, when interested in continuous treatments, a proposal is to use a weighted kernel density estimator (wKDE) (Silverman, 1986), in which, similarly to Paper III, the expected value of the wKDE could be decomposed and minimized leading to optimal weights that balance covariates with respect of a continuous treatment.

- *Constrained optimization for robust causal inference.* In the past few years, several methods have been proposed to stabilize probability weights when the positivity assumption is practically violated, which are embedded within the framework of constrained optimization. Among others, Zubizarreta (2015) suggested minimizing the variance of the weights while balancing covariates. Athey et al. (2016) introduced a more general estimator in case of high dimensions. Kallus (2016) proposed to use a set of probability weights that optimizes covariate balance while regularizing the resulting weights. Future research may evaluate and compare these methods in a comprehensive simulation study, propose extensions and novel formulations of some of the aforementioned methods and provide practical guidelines on the choice of the most suitable method when estimating causal effects in case of violations of the positivity assumption.

- *Optimal covariate balancing for high-dimensional causal inference.* Several methods have been developed to estimate causal effects from non-experimental data under the assumptions of unconfoundedness and covariate overlap, also known as positivity assumption. Popular examples include regression, inverse probability weighting, matching, and covariate balancing. These methods aim at balancing the distributions of the observed features across treated and control groups. In practice, to make the causal assumption of unconfoundedness plausible, investigators include a substantial number of features. However, in these settings, the assumption of covariate overlap is more difficult to satisfy. As a consequence, recently, there has been considerable interest in extending the earlier literature on causal inference from non-experimental data to high-dimensional settings. Future research may focus on developing novel methods that balance covariates when using machine learning to estimate high-dimensional nuisances parameters, which can compensate for the bias of regularized regression adjustments, be a more robust alternative to inverse probability weighting, and extend the proposed method to longitudinal settings where the estimation of the causal effect of a time-varying treatment is of interest. Furthermore, future research may explore the implications of overlap in high-dimensional non-experimental studies and investigate potential extensions to high-dimensional censored data.

References

Athey, S., G. W. Imbens, and S. Wager. 2016. Approximate Residual Balancing: De-Biased Inference of Average Treatment Effects in High Dimensions. *ArXiv e-prints* .

Athey, S., G. W. Imbens, S. Wager, et al. 2016. Efficient inference of average treatment effects in high dimensions via approximate residual balancing. Technical report.

Bang, H., and J. M. Robins. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics* 61(4): 962–973.

Basu, D. 2011. An Essay on the Logical Foundations of Survey Sampling, Part One. In *Selected Works of Debabrata Basu*, ed. A. DasGupta, 167–206. Selected Works in Probability and Statistics, Springer New York.

Beaumont, J.-F. 2008. A new approach to weighting and inference in sample surveys. *Biometrika* 95(3): 539–553. URL `http://biomet.oxfordjournals.org/content/95/3/539`

Beaumont, J.-F., D. Haziza, and A. Ruiz-Gazen. 2013. A unified approach to robust estimation in finite population sampling. *Biometrika* 100(3): 555–569. URL `http://biomet.oxfordjournals.org/content/100/3/555`

Blackwell, M. 2013. A framework for dynamic causal inference in political science. *American Journal of Political Science* 57(2): 504–520.

Bottai, M., and J. Zhang. 2010. Laplace regression with censored data. *Biometrical Journal* 52(4): 487–503.

Cassel, C. M., C. E. Särndal, and J. H. Wretman. 1976. Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika* 63(3): 615–620.

Chakraborty, B., S. Murphy, and V. Strecher. 2010. Inference for non-regular parameters in optimal dynamic treatment regimes. *Statistical methods in medical research* 19(3): 317–343.

Chakraborty, B., and S. A. Murphy. 2014. Dynamic treatment regimes. *Annual review of statistics and its application* 1: 447–464.

Chan, K. C. G., S. C. P. Yam, and Z. Zhang. 2016. Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78(3): 673–700.

Cole, S. R., and M. A. Hernán. 2008. Constructing inverse probability weights for marginal structural models. *American journal of epidemiology* 168(6): 656–664.

Cox, B., and D. McGrath. 1981. An Examination of the Effect of Sample Weight Truncation on the Mean Square Error of Survey Estimates. *Paper Presented at the 1981 Biometric Society ENAR Meeting* Richmond, VA, U.S.A.

Daniel, R., S. Cousens, B. De Stavola, M. Kenward, and J. Sterne. 2013. Methods for dealing with time-dependent confounding. *Statistics in medicine* 32(9): 1584–1618.

Deville, J.-C., and C.-E. Särndal. 1992. Calibration estimators in survey sampling. *Journal of the American statistical Association* 87(418): 376–382.

Elliot, M., and R. Little. 2000. Model-based alternatives to trimming survey weights. *Journal of Official Statistics* 16(3): 191.

Elliott, M. R. 2009. Model Averaging Methods for Weight Trimming in Generalized Linear Regression Models. *Journal of official statistics* 25(1): 1–20. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3530169/`

Foreman, E., and K. R. Brewer. 1971. The efficient use of supplementary information in standard sampling procedures. *Journal of the Royal Statistical Society. Series B (Methodological)* 391–400.

Freedman, D. A. 2006. On the so-called "Huber sandwich estimator" and "robust standard errors". *The American Statistician* 60(4): 299–302.

Friedman, J., T. Hastie, and R. Tibshirani. 2001. *The elements of statistical learning*, vol. 1. Springer series in statistics New York.

Griva, I., S. G. Nash, and A. Sofer. 2009. *Linear and nonlinear optimization*, vol. 108. Siam.

Gruber, S., R. W. Logan, I. Jarrín, S. Monge, and M. A. Hernán. 2015. Ensemble learning of inverse probability weights for marginal structural modeling in large observational datasets. *Statistics in medicine* 34(1): 106–117.

Hainmueller, J. 2012. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis* 20(1): 25–46.

Hernán, M. Á., B. Brumback, and J. M. Robins. 2000. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men.

Hernán, M. A., B. Brumback, and J. M. Robins. 2001. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association* 96(454): 440–448.

Hernan, M. A., and J. M. Robins. 2010. *Causal inference.* CRC Boca Raton, FL:.

Hirshberg, D. A., and S. Wager. 2017. Balancing Out Regression Error: Efficient Treatment Effect Estimation without Smooth Propensities. *ArXiv e-prints* .

HIV-Causal Collaboration, et al. 2010. The effect of combined antiretroviral therapy on the overall mortality of HIV-infected individuals. *AIDS (London, England)* 24(1): 123.

———. 2011. When to initiate combined antiretroviral therapy to reduce mortality and AIDS-defining illness in HIV-infected persons in developed countries: an observational study. *Annals of Internal Medicine* 154(8): 509.

Hofmann, T., B. Schölkopf, and A. J. Smola. 2008. Kernel methods in machine learning. *The annals of statistics* 1171–1220.

Horvitz, D. G., and D. J. Thompson. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association* 47(260): 663–685.

Hulliger, B. 1995. Outlier Robust Horvitz-Thompson Estimators 21(1): 79–87.

Imai, K., and M. Ratkovic. 2014. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76(1): 243–263.

———. 2015. Robust estimation of inverse probability weights for marginal structural models. *Journal of the American Statistical Association* 110(511): 1013–1023.

Imbens, G. W., and D. B. Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences.* Cambridge University Press.

Kallus, N. 2016. Generalized Optimal Matching Methods for Causal Inference. *ArXiv e-prints* .

Kang, J. D., and J. L. Schafer. 2007a. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science* 523–539.

Kang, J. D. Y., and J. L. Schafer. 2007b. Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statist. Sci.* 22(4): 523–539. URL `http://dx.doi.org/10.1214/07-STS227`

Karim, M. E., J. Petkau, P. Gustafson, H. Tremlett, and T. B. S. Group. 2017. On the application of statistical learning approaches to construct inverse probability weights in marginal structural cox models: Hedging against weight-model misspecification. *Communications in Statistics-Simulation and Computation* 1–30.

Karim, M. E., and R. W. Platt. 2017. Estimating inverse probability weights using super learner when weight-model specification is unknown in a marginal structural Cox model context. *Statistics in Medicine* 36(13): 2032–2047.

Koenker, R. 2011. A note on Laplace regression with censored data. *Biometric Journal* 53(5): 855–860.

Kokic, P., and P. Bell. 1994. Optimal winsorizing cutoffs for a stratified finite population estimator. *Journal of Official Statistics* 10(4): 419.

Kuhn, H. W. 1955. The Hungarian method for the assignment problem. *Naval Research Logistics (NRL)* 2(1-2): 83–97.

Lee, B. K., J. Lessler, and E. A. Stuart. 2010. Improving propensity score weighting using machine learning. *Statistics in medicine* 29(3): 337–346.

Lefebvre, G., J. A. Delaney, and R. W. Platt. 2008. Impact of mis-specification of the treatment model on estimates from a marginal structural model. *Statistics in medicine* 27(18): 3629–3642.

Li, F., K. L. Morgan, and A. M. Zaslavsky. 2017. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association* 1–11.

Li, Y. P., K. J. Propert, and P. R. Rosenbaum. 2001. Balanced risk set matching. *Journal of the American Statistical Association* 96(455): 870–882.

Little, R. J., and D. B. Rubin. 2014. *Statistical analysis with missing data*, vol. 333. John Wiley & Sons.

Lodi, S., D. Costagliola, C. Sabin, J. del Amo, R. Logan, S. Abgrall, P. Reiss, A. van Sighem, S. Jose, J.-r. Blanco, et al. 2017. Effect of immediate initiation of antiretroviral treatment in HIV-positive individuals aged 50 years or older. *Jaids Journal of Acquired Immune Deficiency Syndromes* .

Lu, B. 2005. Propensity score matching with time-dependent covariates. *Biometrics* 61(3): 721–728.

Luenberger, D. G., and Y. Ye. 2015. Linear and Nonlinear Programming .

Lumley, T., P. A. Shaw, and J. Y. Dai. 2011. Connections between survey calibration estimators and semiparametric models for incomplete data. *International Statistical Review* 79(2): 200–220.

McAfee, A. J., E. M. McSorley, G. J. Cuskelly, B. W. Moss, J. M. Wallace, M. P. Bonham, and A. M. Fearon. 2010. Red meat consumption: An overview of the risks and benefits. *Meat science* 84(1): 1–13.

Pfeffermann, D. 1993. The role of sampling weights when modeling survey data. *International Statistical Review/Revue Internationale de Statistique* 317–337.

Pfeffermann, D., and M. Sverchkov. 1999. Parametric and Semi-Parametric Estimation of Regression Models Fitted to Survey Data. *Sankhya: The Indian Journal of Statistics, Series B (1960-2002)* 61(1): 166–186. URL `http://www.jstor.org/stable/25053074`

Potter, F. 1990. A study of procedures to identify and trim extreme sampling weights. In *Proceedings of the American Statistical Association, Section on Survey Research Methods*, vol. 225230.

Rao, J. N. K. 1966. Alternative Estimators in PPS Sampling for Multiple Characteristics. *Sankhya: The Indian Journal of Statistics, Series A (1961-2002)* 28(1): 47–60. URL `http://www.jstor.org/stable/25049398`

Rasmussen, C. E. 2004. Gaussian processes in machine learning. In *Advanced lectures on machine learning*, 63–71. Springer.

Rivest, L.-P., D. Hurtubise, and Statistics Canada. 1995. On Searls' winsorized mean for skewed populations. In *Survey methodology*, 107–116.

Robins, J. 1986. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling* 7(9-12): 1393–1512.

Robins, J., M. Sued, Q. Lei-Gomez, and A. Rotnitzky. 2007. Comment: Performance of Double-Robust Estimators When "Inverse Probability" Weights Are Highly Variable. *Statist. Sci.* 22(4): 544–559. URL `http://dx.doi.org/10.1214/07-STS227D`

Robins, J. M. 2000. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical models in epidemiology, the environment, and clinical trials*, 95–133. Springer.

Robins, J. M., S. Greenland, and F.-C. Hu. 1999. Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome. *Journal of the American Statistical Association* 94(447): 687–700.

Robins, J. M., A. Rotnitzky, and L. P. Zhao. 1994. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association* 89(427): 846–866.

———. 1995. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the american statistical association* 90(429): 106–121.

Rosenbaum, P. R. 1989. Optimal matching for observational studies. *Journal of the American Statistical Association* 84(408): 1024–1032.

———. 2002. Observational studies. In *Observational studies*, 1–17. Springer.

Rosenbaum, P. R., and D. B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1): 41–55.

Santacatterina, M., R. Bellocco, A. Sönneborg, A. Ekström, and M. Bottai. 2017. Optimal probability weights for estimating causal effects of time-varying treatment with marginal structural Cox models.

Särndal, C. E. 1980. On π-inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika* 67(3): 639–650.

Scharfstein, D. O., A. Rotnitzky, and J. M. Robins. 1999. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* 94(448): 1096–1120.

Schölkopf, B., and A. J. Smola. 2002. *Learning with kernels: support vector machines, regularization, optimization, and beyond.* MIT press.

Seaman, S. R., and I. R. White. 2013. Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research* 22(3): 278–295.

Silverman, B. W. 1986. *Density estimation for statistics and data analysis*, vol. 26. CRC press.

Stuart, E. A., S. R. Cole, C. P. Bradshaw, and P. J. Leaf. 2011a. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 174(2): 369–386.

———. 2011b. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society Series A* 174(2): 369–386. URL `http://onlinelibrary.wiley.com/doi/10.1111/j.1467-985X.2010.00673.x/abstract`

Tsiatis, A. 2007. *Semiparametric theory and missing data.* Springer Science & Business Media.

Valliant, R., J. A. Dever, and F. Kreuter. 2013. *Practical tools for designing and weighting survey samples.* Springer.

Wong, R. K., and K. C. G. Chan. 2017. Kernel-based covariate functional balancing for observational studies. *Biometrika* .

Xiao, Y., E. E. Moodie, and M. Abrahamowicz. 2013. Comparison of approaches to weight truncation for marginal structural Cox models. *Epidemiologic Methods* 2(1): 1–20.

Zhao, Q. 2016. Covariate Balancing Propensity Score by Tailored Loss Functions. *ArXiv e-prints* .

Zhao, Y.-Q., D. Zeng, E. B. Laber, and M. R. Kosorok. 2015. New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association* 110(510): 583–598.

Zubizarreta, J. R. 2015. Stable Weights that Balance Covariates for Estimation With Incomplete Outcome Data. *Journal of the American Statistical Association* 110(511): 910–922. URL `http://dx.doi.org/10.1080/01621459.2015.1023805`

Acknowledgements

There are many people that I would like to thank for their contributions to this thesis, and for their support and encouragement during these years.

Thanks to **Matteo Bottai**, my main supervisor, for giving me the opportunity of doing this Ph.D. and for everything you have taught me. A special thanks to my co-supervisors **Anders Sönnerborg**, **Anna Mia Ekström**, and **Rino Bellocco** for sharing with me all your scientific experience and for all the support you gave me.

I wish to express my sincere gratitude to **Rino Bellocco**. It is only because of you that I had the opportunity to start this incredible journey into research.

A particular thanks to **Celia Garcia Pareja** and **Pol Del Aguila Pla** for the invaluable scientific and psychological support you guys gave me during these years. Thank you for being a friend, a psychologist and a supervisor at the same time.

I wish to thank **Giorgio Raccanelli** for being one of the few people who believed in me when I was younger.

Thank you very much, **Lena Palmberg**, for all your help and understanding during this past year.

Thanks to all my colleagues at the Unit of Biostatistics and IMM: **Celia Garcia Pareja**, **Andrea Discacciati**, **Paolo Frumento**, **Federica Laguzzi**, **Xin Fang**, **Erin Gabriel**, **Michael Sachs**, **Jonas Höijer**, **Ulf Hammar**, **Yang Cao**, and **Daniel Olsson**.

I wish to thank **Fabio Giudici** for always being a true friend, a fellow traveller and the greatest dragon I have ever met in my life, **Piero Gasparotto** and **Alberto Testolin** for teaching me how to be grateful for the simple things and fully appreciate life, and **Hoyin Lam** for being one of the brightest researchers and people I have met in these past years.

A special thanks to my family, **Giliana Wally**, **Adriano**, **Giulia**, **Anita**, **Vasco**, **Antonietta**, **Silvano** and **Annamaria** for all the values you taught me and for helping me become who I am.

Thank you, **Misterpom Little Jewels Princess Cake Tubular Shadow**, also known as **PK**, for making my days brighter with your fluffiness, beauty and ferocious barks.

Above all, I would like to thank my fiancée **Alice**. *Non ci sono parole per descrivere la mia gratitudine nei tuoi confronti. Mi hai seguito, consolato, capito e aiutato durante tutto questo periodo rendendomi ogni giorno sempre più felice. Sei la persona più importante della mia vita e lo sarai per sempre.*

This work was supported by the **Karolinska Institutet**'s funding for doctoral students (KID-funding), and the **ALF** Medel Stockholms Läns Landsting.