Inverse Probability of Treatment Weighting Under Violations of Positivity: Draft 2

Tomas D. Morley

March 4, 2018

1 Abstract

In this thesis we study the performance of the inverse probability of treatment weighted (IPTW) for marginal structural models under violations of the positivity assumption. To our knowledge the performance of MSMs under violations of this assumption have not been systematically studied in the litrature. We employ the simulation algorithm of Havercroft and Didelez (2012) to generate data from a typical longitudinal setting. By exploiting the design of this algorithm it is possible to introduce positivity violations that are propagated through patient histories and study the effect of these violations on the ability of a marginal structural model to recover the true parameters. Our results suggest that even small violations of the positivity assumption can have a large impact on the performance of the estimation.

2 Acknowledgements

Contents

1 Abstract **2**

2 Acknowledgements **3**

3 Section 1: Introduction **5**
- 3.0.1 Directed Acyclic Graphs: graphical representations of causality 6
- 3.0.2 Notation 6
- 3.1 Time Dependent Confounding 6
  - 3.1.1 Marginal structural models 7
- 3.2 Inverse Probability Weighting 10
- 3.3 Simulating from a MSM 11
  - 3.3.1 collapsibility 13
- 3.4 Positivity 13
- 3.5 Static vs. Dynamic Strategies 15
- 3.6 Related work 15

4 Chapter 2 **18**
- 4.1 Simulating from a static MSM 18
- 4.2 Data Structure 18
- 4.3 Simulation Algorithm 19
  - 4.3.1 Algorithm 19
  - 4.3.2 Discussion of how algorithm works 21
- 4.4 Constructing IPT weights 21
- 4.5 Simulation Set-up 22
- 4.6 Results 22

5 Dynamic Case **22**
- 5.1 The problem of simulating from a MSM under a dynamic strategy 22

6 Violations of Positivity **22**
- 6.1 Extended discussion of algorithm linking to positivity 22

7 Application **23**

8 Discussion and Conclusion **23**
- 8.1 Limitations 23

3 Section 1: Introduction

Marginal structural models (MSMs) are a popular class of models for performing causal inference in the presence of time dependent confounders. These models have an important application in areas of research such as epidemiology, social sciences and economics where randomised trials are prohibited by ethical or financial considerations, and hence confounding cannot be ruled out by randomization. Under these circumstances confounding can obscure the causal effect of treatment on outcome. An example of this, common in epidemiological studies, occurs when prognostic variables inform treatment decisions while at same time being predictors of the outcome of interest. In a longitudinal setting this is further complicated when the confounder itself is determined by earlier treatment. One consequence is that regression adjustment methods do not control for confounding in the longitudinal case and other techniques are required. A second consequence is that simulating data from a specific marginal structural models is more challenging when the data is to exhibit time dependent confounding.

The Inverse probability of treatment weighting (IPTW) estimator is a technique that has been applied to censoring, missing data and survey design problems. The central idea is that by weighting the observed data, a pseudo population is constructed in which tratemnt is assigned at random. Subsequent analysis where we ignore the confounder is then possible. which inference on the target population can be achieved. For example, when there is missing data weights can be used to create a pseudo-population in which there is no missingness. In the context of MSMs, the IPT weights relate to a pseudo-population in which there is no longer any confounding between the confounder and treatment and causal inferences can be made.

Underlying IPTW method for estimating MSMs are four assumptions: 1) consistency 2) exchangeability 3) positivity 4) and correct model specification. Exchangeability, also known as the no unmeasured confounding assumption, is closely linked to causality?? Several studies have considered violations of exchangeability and corrected model specification. Positivity has received less attention because in typical observational study positivity violations are not suspected explain why. In the clinical context that we consider, protocols threaten to violate the positivity assumption and we investigate whether MSMs are robust against positivity. The focus of this thesis will be on violations of the positivity assumption. Positivity means that within every strata spanned by the confounders, there must be a positive probability of patients being exposed or unexposed to treatment. For example, in a medical context, if treatment protocols demand that treatment is initiated whenever a prognostic variable falls below a pre-defined threshold, there will only be exposed and no unexposed patients in this strata of the confounding prognostic variable. make decisions based on protocols positivity can be. In the absence of structural positivuty violations, there is always the threat that random zeroes arise in some strata of the confounder especially when the sample size is small or the number of confounding variables is large. In each case the sparsity of data within the strata of the confounder results in a high chance that positivity is violated. Positivity violations increase the bias and variance of estimates of the causal effect but the extent of the damage is not well known. The central aim of this thesis will be to investigate positivity violations when fitting MSMs to longitudinal data. To our knowledge positivity violations have not been systematically studied in the literature from a simulation point of view. We quantify the bias and variance introduced due to positivity violations and hope to provide practical advice to researchers tempted to fit MSMs to overcome confounding without realising the potential consequences of positivity violations in their data.

Throughout this thesis we focus on clinical applications as examples. In the literature on marginal structural models the causal effect of Zidovudine on the survival of HIV positive men is often cited as an example. In this example a patients white blood cell (CD4) count is a prognostic

variable that influences a doctor's decision to initiate treatment while at the same time being a predictor of survival. As a result CD4 count is a confounder. In the longitudinal setting previous treatments influence CD4 count. As such studies often depend on protocols which means that poistivity in some levels of the confounder make this a suitable example for our purposes.

The structure of this thesis is as follows. In section 2 of part 1, the model considered in this thesis and its important aspects are explained. In part 2 simulating from this statistical model is discussed in detail. In part 3 the model under dynamic strategies is considered and comparisons are drawn with the static case. In part 4 we entertain violations of positivity in the data, this section represents the novelty in this thesis. Part 5 conducts a simulation study. Part 6 includes a discussion, conclusions and suggestions for future work.

Cole et al have discussed the role of positivity underlying the IPTW method but a limitation of these paper is that they lack a formal exploration through simulation of the effects this can have. Here we explore formally different levels of the violation and offer practical advice to researchers who suspect this as an issue.

3.0.1 Directed Acyclic Graphs: graphical representations of causality

In this thesis we consider simulating from marginal structural models (MSMs), a class of models that can be used to estimate causal effects from observational data under time dependent confounding. Specifically, we consider the problem, common in medical research, of investigating the relationship between a binary treatment variable A and an outcome of interest Y in the presence of a covariate L. One way of expressing the relationship between these three variables is through a graph such as the one shown in the left hand size of figure 1.

A graph consists of a finite set of vertices ν and a set of edges ϵ. The vertices of a graph correspond to a collection of random variables which follow a joint probability distribution $P(\nu)$. Edges in ϵ consist of pairs of distinct vertices and denote a certain relationship that holds between the variables Pearl (2009). The direction of the relationship is denoted by an arrow and in this thesis we consider only acyclic graphs which means that the relationship between two variables only proceeds in one direction, there are no feedback loops or mutual causation. The resulting directed acyclic graph (DAG) $G = (\nu, \epsilon)$ shown in figure 1 can be represeted as a set of connections $(A, Y), (L, Y)$ where direction is from the first element of each pair to the second element.

3.0.2 Notation

3.1 Time Dependent Confounding

The single time case of confounding is represented in figure 1 where L affects both A and Y. A model of Y in terms of A such as a regression model where Y is regressed against A will not resolve the correct causal parameters because A is confounded by L. If A and L are independent then this is not an issuethat ignores L will notIf we ignore L from the analysis and model Y in terms of A we will not get the true c The right hand side of figure 1 represents the confounding case where the DAG G has been ammended to include a relationship from L to A $G = (A, Y), (L, Y), (L, A)$. This confounding relationship precludes causal inference because we cannot discern the relationship between treatment A on outcome Y when L is a common cause of both. If L is the only confounder and there are no unmeasured confounders, then it is possible to correctly estimate the causal relationship between A and Y by adjusting for the measured confounders.

time dependent confounding in the set-up we consider, will always introduce bias. In the time dependent case, we will therefore always have bias. And IPT weighting a means of avoiding this.

The time dependent case adds another complication which is a further relationship between previous treatment A_{t-1} and L_t as shown in figure 2.

Talk about the minimum required to deal with the time dependent case. Data on all time independent and time dependent variables. correct model specification. Explain why time dependent as opposed to time independent (baseline - do not change over time) variables are more of a problem.

By blocking the pathway between A -> L -> Y at L, we get distorted estimates of the effect of treatment on outcome.

In the one-shot case the outcome Y depends on the treatment decision, measured covariates and unmeasured covariates. In the longitudinal case, the outcome depends on the histories of the treatment and covariates. One complication is that current values of the covariate may depend on previous treatments, and previous treatments may, in turn, depend on previous covariates. If treatment is succesful this will inform future treatments and also affect the outcome of interest Y. While in the one-shot case in figure 1 regression adjustment using

Allow for the joint determination of outcomes and treatment status or omitted variables related to both treatment status and outcomes (Angrist 2001).

Similarily, if treatment is succesful this will affect both the outcome and subsequent treatments. Analogous to the one-shot case, $P(y \mid \bar{A} = \bar{a}) \neq P(y \mid do(\bar{A} = \bar{a}))$. Regression adjustment will always introduce other sources of bias in the time dependent case. As a result regression adjustment does not work. Only the IPTW method adjust for selection bias and confounding.

A covariate L is a confounder if it predicts the event of interest and also predicts subsequent exposure. Explain how this actually happens, as U0 is a common ancestor of A through L and also Y, also that there is selection bias, and L is sufficient to adjust for confounding see Havercroft algorithm code page bottom.

The reason why U is important is because with U, we can obtain any P(Y | do(a))$ using the inverse of U?? We can do this under any counterfactual survival path.

Explain why there will always be selection bias, and hence there will always be a form of confounding in a longitudinal model. Talk about conditioning on Y in a collider setting.

3.1.1 Marginal structural models

The positivity problem is less well known (Cole2009_consistency paper) He points out that consistency and positivity are less well known because there is little written about them and the problems they create. But there is now a lot written on consistency, but still relatively little on positivity.

Specify explictly that marginal structural models are used for estimating average causal effects.

define what a causal effect is and how we go from the individual to the population and that we cannot in general identify individual causal effects due to a missing data problem. Let the assumptions behind this extend from the meaning of causal effects, for example the exchangeability assumption.

Marginal structural models are a class of models which can be used to estimate causal effects in the presence of time dependent confounding. Specifically we are interested in the causal effect of a treatment A on an outcome Y. A causal model differs from an associative model and associative models like regression models will not have a causal interpretation when time dependent confounding is present, nor will naive methods of adjustment result in a causal interpretation. When time dependent confounding is present, associative models such as regression models do not have a causal effect.

- independence between treatment and prognostic factors/covariates.
- comparison with the random treatment paradigm.
- If treatment at each time t is assigned completely at random then treatment will be independent of both measured and unmeasured confounding
- At the individual level we cannot establish causal effects because of a missing data problem. Namely that we do not see the counterfactual outcomes. Link in literature about all problems being missing data problems.
- The missing data idea could be shown with an example table of results of people under treatment and not under treatment.

They model the marginal distribution of counterfactual variables over any covariates and are referred to as structural because in econometric and social sciences literature the term structural is often used in place of causal (hernan, brumback and robins).

An oft cited observational outcome in statistics without a causal relationship is umbrella sales and the probability of rain. A counterfactual outcome is broader, it encompasses not only the outcome when umbrella sales are purchased in high volumes but also when they are not. Intuitvely the connection between counterfactuals and A counterfactual variable for Y is the random variable consisting of a subjects outcome under a A regardless of whether the subject actually received this treatment or not. Juxtaposed to an observational outcome, a counterfactual outcome ... correspondingly, the probability recorded for a subject. In the observational case, given the treatment and . Within the context of a MSM, the focus is the causal relationship that exists between treatment and outcome. The following from Pearl (2009) defines a causal relationship.

Definition 1 (Def 3.2.1 from Pearl (2009), abridged) Given two disjoint sets of variables, A and Y, the causal effect of A on Y, denoted as $P(y \mid do(A = a))$, is a function from X to the space of probability distributions on Y. For each realization a of A, $P(y \mid do(A = a))$ gives the probability that $Y = y$ induced by deleting from the model all equations corresponding to variables in A and substituting a into the remaining equations.

Associative models like regression models model the conditional distribution $P(Y \mid A = a, L = l)$

8. Both the structural and exposure models must be correctly specified.

Need a better description of do notation and how it differs to conditional case and how it relates to counterfactuals

See Robins etal 2000 for explanantion of why they are called marginal

explain counterfactual and how we model it (this is what is missing and need to connect to graph)

marginal because they model the marginal distribution of the counterfactual variables. Not marginal over covariates??

$E(Y \mid do(a)) = \beta_0 + \beta_1 a$ contains no l term.

Non collapsible models and simulation

Have been used for missing data problems. see pp.442 of Hernan, Brumback, Robins 2001 for a list of papers linked to this

structural is only in the MSM name because structural is used in econometrics and social science. Structural here only means that it is a causal model.

Counterfactuals, breeze through this ...

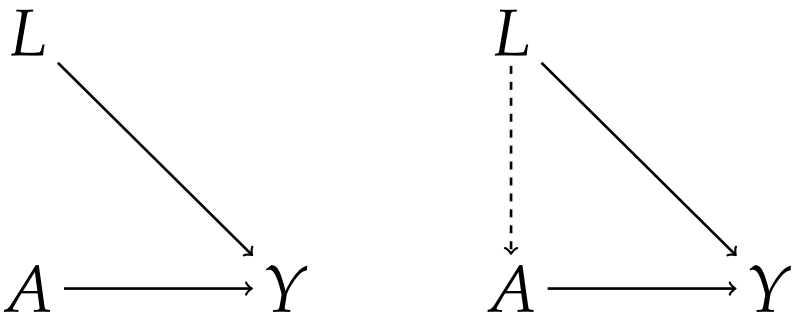

Figure 1 DAG

A MSM When a relationship exists between L and A as in the right hand side of figure 1 then there is a confounding relationship and associative models will not have a causual interpretation. The following definition from Pearl (2009) defines what is meant by a causal effect:

An example of this definition involves deleting the link between L and A in the diagram.

An associative model like a regression model.

When treatments are randomised this is not a problem This might arise, for example, when the decision to instigate treatment is based on a measured covariate L. Contrast this to randomized controlled trial, where due to the randomization process treatment will be independent of any measured or unmeasured covariates. The randomized controlled trial does not often arise in epidemiology, social sciences or econometrics due to prohibitive ethical or financial concerns. As a result other methods are sought in order to uncover the causal effects. Marginal structural models are a class of models which make it possible to estimate the causal effect of A on Y in the presence of confounders.

- still have unmeasured confounders of course. Or model mispecification. Either of these mean that results no longer hold.

A model that parameterises $P(Y \mid do(A = a))$ is called a marginal structural model (MSM) as it is marginal over any covariates and structural in the sense that it represents an interventional rather than observational model. The distinction between an interventional and observational model is expressed in definition 1.1 through the notation $do(a)$ in $P(y \mid do(x))$, as opposed to the observational case $P(y \mid A = a)$.

marginal structural models are usually for counterfactuals -> example -> straight through to the IPW not too much on counterfactuals.

Randomization ensures that missing values occur by chance. So the counterfactual values that we don't see for some observations are missing randomly and not due to confounding through a covariate.

An example of this distinction is the effect of Zidovudine on the survival of HIV-positive men. Of interest is the effect that Zidovudine has on survival. This is what we want to estimate and it is captured in the DAG in figure 1 as the edge connecting A and Y.

importantly, changing the relationship between L and A, won;t change the relationship between L and Y. This means that an intervention in A does not affect the relationship between L and Y. So we remove the link between L and A and assign to A the value of treatment on or off. Once we place the patient on treatment, regardless of the relationship which had existed before hand between the covariate and treatment, a new relationship between A and Y exists in which the covariate has no say.

This distinction can be understood more clearly when comparing randomized controlled trials in medical resarch and observational epidemiological research. In randomised trials the In the presence of confounding $P(y \mid A = a)$ will not represent the true causal effect of A on Y why?. The interventional case $P(y \mid do(A = a))$, on the otherhand, $\neq P(y \mid do(x))$ with the result that a model which is conditional on A will not represent the true causal effect of A on Y. In the former case the system is changed such that variable A takes on a particular value or history, the result represents the system under the intervention $do(A = a)$ (pearl 2010). On the other hand the first method represents the situation where the model is unchanged and A is observed to be a. This distinction will be particularly important when considering static and dynamic stratgeies in subsequent sections.

From the point of view of simulating from a marginal structural model, In a nutshell, we need a simulation algorithm that allows us to simulate discrete time longitudinal data in which A affects Y, and so does L and L affects A. The most direct way of doing this would be the red line in the DAG. But this won't do in our case. The Havercroft and Didelez algorithm is useful because it allows us to break the relationship between L and A. We need to do this in such a way that we do not change the relationship between Y and L, although we are not so interested in this relationship because it is the least interesting of the relationships in the diagram.

3.2 Inverse Probability Weighting

Inverse probability of treatment weighting is a technique that re-weights subject observations to a population where assignment of treatment is at random. An early example of this technique is the **?** weighted estimator of the mean. In the context of marginal structural models, a weight is calculated for each subject which can be thought of informally as the inverse of the probability that a subject receives their own treatment Robins et al. (2000). The result of applying these weights is to re-weight the data to create a pseudo-population in which treatment is independent of measured confounders Cole and Hernán (2008). The resulting pseudo population no longer has a time dependent confounding problem and the causal relationship between A and Y can be estimated using naive methods like regression without adjusting for confounding because it is no longer present. Crucially, in the pseudo population the counterfactual probabilities are the same as in the true study population so that the causal RD, RR or OR are the same in both populations Robins et al. (2000).

$$w_{t,i} = \frac{1}{\prod_{\tau=0}^{t} p_\tau(A_{\tau,i} \mid \bar{A}_{\tau-1,i}, \bar{L}_{\tau,i})}$$

stabilized weights

$$sw_{it} = \frac{\prod_{\tau=0}^{t} p_\tau(A_{\tau,i} \mid \bar{A}_{\tau-1,i})}{\prod_{\tau=0}^{t} p_\tau(A_{\tau,i} \mid \bar{A}_{\tau-1,i}, \bar{L}_{\tau,i})}$$

The use of IPTW is valid under the four assumptions of consistency, exchangeability, positivity and no misspecification of the model Cole and Hernán (2008).

Informally a patients weight through visit k is proportional to the inverse of the probability of having her own exposure history through visit k (Cole and Hernan 2008)

In previous simulation studies unstabilized weights show substantial increase in SEs

Explain why this method does appropriately adjust for **measured time varying** confounders affected by prioir exposure.

The weight is informally proportional to the participants probability of receiving her own exposure history

As these weights have high instability we need to stabilize them. The unstabilized weights can be driven by only a small number of observations. Under time dependent confounding it may still be possible to recover the causal effect of A on Y by the method of Inverse Probability of Treatment (IPT) weihting. How does this work?

- true weights are unknown but can be estimated from the data.
- Robins(2000) - when there are no unmeasured confounders, we can control for confounding using weights
- weighting adjusts for confounding and selection bias due to mesured time varying covariates affected by prior exposure.
- A_t is no longer affected by L_t, and crucially the causal effect of $\bar{A}$ on Y remains unchanged
- quantify the degree to which treatment is statistically non-exogenous through month t

lack of adjustment for L precludes unbiased estimation. This is because it introduces selection bias.

Problem is that the weights will have very high variability, so we need to stabilize them otherwise they will affect the estimates. Explain all the little details about why it becomes more variable etc.

No matter whether we use stabilized or unstabilized weights, under positivity violations the weights will be undefined.

Conditions under which IPTW work are largely untestable (westreich 2012)

Be more specific about what is contained in the weights. The denominator depends on the measured confounders L the numerator does not, but could contain baseline coefficients in the numerator to help stabilize the weights.

weighted regression and MSM are equivalent.

L0 and baseline values - make sure that B = baseline and L0 are either together, or have them clearly split. V is a variable predictive of A but not of Y - hence not a confounder No rebalancing the baseline covariates like gender or age (B), so if we want to see the effect of groups of basline If we do not condition on baselines like age then we break the link between age and treatment. So baselines must be in the condition.

3.3 Simulating from a MSM

In this section we discuss the simulation algorithm of Havercroft and Didelez (2012). which is later used to explore the affects of positivity violations on the IPTW method. In particular we describe how this algorithm relates to the time dependent confoudning described above, and how it allows us to capture the observational nature of the data while still permitting simulating from a given MSM.

How does selection bias relate to U?

Explain why the algoritm allows positivity to be propagated throught the patients history.

The simulating procedure needs to allow us to specifiy a particular MSM, while at the same time capturing the characteristics of observational data.

Problem statement: Simulating from a given marginal structural model is challenging because a direct relationship between L and Y would ...

Problem is that conditional distributions one might use to draw the simulated data are not compatible with the desired properties of the marginal model.

1. derive relationship between MSM and DAG and the correct conditional distributions. Follows from truncated factorisation why we can et P(Y | do(a))

2. simulating from a chosen MSM
3. creating a confounding relationship
4. only have the data and observed outcomes, not the counterfactuals.
5. Fixing relationships.
6. think of this process as if we had fixed a treatment vector in advance. consistency assumption.
7. Relationship between A and L should be observational and a confounding relationship
8. U is breaking the relationship between L and Y, important for positvity.
9. Don't care about the relationship between L and Y
10. HD 2012, with Pearl and truncated factorization formula, show that it is possible to link the counterfactual represented by $P(Y \mid do(a))$ to observattional data generated in an observational way. But the problem arises when the model is non-collapsible or non-linear.

In the one shot case we set A = 1/0 because we are interested in the outcome under either of these treatment scenarios. In the time dependent case, A is a vector of 0s and 1s and we want to pretend that we decide in advance that the whole vector A is specified. But A and L have a complex interplay in an observational setting. So we want to pretend that A (a vector of 1s and 0s) is set in advance but at the same time have the observational structure for A and L.

The relationship between Y and L is then dependent on A. There is no relationship between A and U because of the set-up in the DAG. The variable L blocks this relationship.

In their paper Havercroft and Didelez (2012) develop an algorithm that allows simulating data that corresponds to a particular parameterisation of an MSM. This algorithm provides the bedrock of the simulation structure considered in this thesis. Figure 1 represents the system under consideration. The DAG in figure 1 represents the one-shot non-longitudinal case. Factorising the joint distributions of the variables in figure 1 yields

$$P(U,\ L,\ W,\ A,\ Y) = P(W)P(U)P(W)P(L \mid U)P(A \mid L, W)P(Y \mid U, A)$$

Where, following definition 1.1 we delete $P(A \mid L, W)$, a probability function corresponding to A, and replace $A = a$ in all remaining functions

$$P(U, L, W, Y \mid do(A = a)) = \begin{cases} P(U)P(L \mid U)P(Y \mid U, A = a) & \text{if } A = a \\ 0 & \text{if } A \neq a \end{cases}$$

The goal is to simulate from a particular MSM. This means parameterising $P(Y \mid do(A = a))$. Applying the law of total probability over W, U and L yields

$$P(Y \mid do(A = a) = \sum_{w,u,l} P(W)P(U)P(L \mid U)P(Y \mid U, L, A = a) = \sum_{u,l} P(U)P(L \mid U)P(Y \mid U, L, A = a)$$

Making use of the fact that $P(L, U) = P(L \mid U)P(U) = P(U \mid L)P(L)$ and summing over either W and U or W and L yields

$$P(Y \mid do(A = a) = \sum_l P(Y \mid L, A = a)P(L) = \sum_u P(Y \mid U, A = a)P(U))$$

If we can find suitable forms for either $P(Y \mid L, A = a)$ and $P(L)$ or $P(Y \mid U, A = a)$ and $P(U)$ that correspond to the MSM $P(Y \mid do(A = a)$, then, given suitable values for A, L, U it will be possible to simulate from the chosen MSM.

use the phrase, "in words ...." to make something more understandable.

Choosing a functional form for

$$P(Y \mid do(A = a)$$

depends on convenience. We need a functional form that can be easily represented by $P(Y \mid L, A = a)P(L)$. non-linear functions will be hard to work into the analysis.

U ~ U[0, 1] is a good choice because we can usethe CDF of Y because U[0, 1] is always between 0 and 1

General health is patient specific but comes from a clear distribution and has a nice medical interpretation. In contrast L would be more difficult to include. It is better as a function of U than a value in of itself.

3.3.1 collapsibility

Collapsibility means there is no incompatibility between the marginal model and the conditional distributions used to simulate the data. Provide example of this. Explain how this affects the simulation algorithm

3.4 Positivity

The final assumption underlying MSMs, and the central topic of this thesis, is the positivity assumption. MSMs are used to estimate average causal effects in the study population, and one must therefore be able to estimate the average causal effect in every subset of the population defined by the confounders Cole and Hernán (2008). The positivity assumption requires that there be exposed and unexposed individuals in every strata of the confounding covariates. For example, when treatment is Zidovudine and CD4 count is the confounder, there must be a positive probability of some patients being exposed and unexposed at every level of CD4 count. Positivity can be expressed formally as $Pr(A = a \mid L) > 0$ for all $a \in A$, which extends straightforwardly to the time dependent case where the positivity assumption must hold at every time step conditonal on previous treatment, time dependent confounders and any baseline covaraiates:

$$Pr(A_{it} = a_{it} \mid L_{it}, A_{i,t-1}, V_{i0}) > 0$$

Models for the risk $P(Y = 1 \mid A = a, L = l)$ are commonly studied in epidemiological applications. Applying basic probability rules reveals that the risk can be re-written with the term $Pr(A = a \mid L = l)$ in the denominator:

$$P(Y = 1 \mid A = a, L = l) = \frac{P(Y = 1, A = a, L = l)}{Pr(A = a, L = l)} = \frac{P(Y = 1, A = a, L = l)}{Pr(A = a \mid L = l)Pr(L = l)}$$

This model is only estimable when $Pr(A = a \mid L = l) \neq 0$. Therefore, when positivity does not hold it is not possible to estimate the model. In the context of MSMs a similar problem emerges. Although weighting via IPTW allows naive estimation of (?) without including the confounders, the weights in (?) involved the term $Pr(A = a \mid L = l)$ in the denominator. This means that the

weights are inestimable whenever positivity is violated. In order to estimate the causal effect of A on Y, weights must be estimable in every subset of the population otherwise the average causal effect in the study population cannot be estimated.

In practice, positivity can arise when random zeroes or structural zeroes are present in some levels of the confounding covariates. Random zeroes arise when, by chance, no individuals or all individuals, receive treatment within a certain strata as defined by the covariates. For example, Cole and Hernán (2008) studies positivity violations in individuals in strata defined by CD4 count and viral load. By increasing the levels of CD4 count the chances of random zeroes also increases and Cole and Hernán (2008) show that the IPT weights rapidly lose their stability with the consequence that causal effects are no longer estimable. Researchers applying IPTW methods must actively check that there are both treated and untreated individuals at every level of their covariates within cells defined by their covariates because parametric methods will smooth over positivity violations and not provide any indication of nonpositivity. Increasingly refined covariates are attractive because they provide better control of confounding, but the point that Cole and Hernán (2008) make is that this control needs to be traded off against increased occurence of random zeroes and subsequent instability of IPT weights.

More relevant to this thesis are violations of the positivity assumption due to structural zeroes. These occur when an individual cannot possibly be treated or if an individual is always treated within some levels of the confounding covariate, as is the case in the clinical protocol example motivating this thesis. Several studies give examples of structural violations of the positivity assumption in epidemiological contexts. In Cole and Hernán (2008) structural zeroes arise when the health effects due to exposure to a chemical are confounded by health status proxied by being at work. If individuals can only be exposed to the chemical at work then all individuals not at work will be unexposed. A second example is liver disease as a contraindication of treatment. If individuals with liver disease cannot be treated then all individuals in the "liver disease = 1" strata will be untreated. In Messer et al. (2010) structural zeroes arise in the context of rates of preterm birth and racial segregation, whereas Cheng et al. (2010) find structural zeroes in the context of fetal position and perinatal outcomes. Our motivating example is most closely related to liver disease as a contraindication, except that the clinical protocols require that patients with low CD4 count always be treated instead of never being treated, as in the case in the liver disease example.

Although in many epidemiological settings the positivity assumption is guaranteed by experimental design, studying positivity violations is relevant because, as our own motivating example and the examples above suggest, structural violations do occur, and random zeroes are always possible especially at finer levels of confounding covariates. Studying the finite sample propoerties of MSMs under violations to positivity is therefore an important issue which is yet to be dealt with systematically in the literature. As Westreich and Cole (2010) points out, positivity violations, positivity violations by a time varying confounder pose an analytic challenge and they suggest g-estimation or g-computation may be a way forward. A good start to dealing with the time varying confounder case is to see how well MSMs work when positivity is violated. This is also a novelty of this thesis.

6. estimated weights with a mean far from one, or very extreme values indicate either nonpositivity or model mispecification of the weight model.
7. It is not always true that we want more finely tuned covariates for confounder control because the bias and variance of the effect estimate may increase with the number of categories. This is similar to the positivity masking example.
8. Our results are equally valid for other circumstances in which positivity may arise.
9. Also think about how the number of categories of exposure increases the chance that one

level of exposure will have a positivity.

10. Westreich and Cole 2010 have suggested that methodological approaches are needed to weigh the resultant biases incurred when trading of confounding and positivity. The framework we use is flexible enough to allow this in a simulation setting.

If the structural bias occurs within levels of a time-dependent confounder then restriction or censoring may lead to bias whether one uses weighting or other methods (Cole and Hernan 2008). In fact, weighted estimates are more sensitive to random zeroes (Cole, Hernan, 2008) Introducing violations of positivity can be achieved by censoring observations.

But to give an intuitive example, think about how it links back to a situation where sicker patients receive treatment compared to others. So in the "sick" strata of the CD4 count **ALL** patients receive treatment which inflates the IPTW. This also affects how we think about the associational versus causal models. The causal effect might be 50/50 but because sicker patients get treatment the mortality ratio in the treated group is likely to be higher.

3.5 Static vs. Dynamic Strategies

So far we have considered static strategies, in this section we desbcribe the differences between static and dynamic strategies. A static stratgey is one where the value where the values that A will take depend on for A is represented as:

Where a(t) = 1 if the strategy specifies that a is to take the value 1 at time t. In contrast, a dynamic regime is any well-specified way of adjusting the choice of the next decision(treatment or dose to administer) in the light of previous information constitutes a dynamic decision (or treatment) strategy (Didelez arxiv).

To our knowledge positivity violations in the dynamic case have not been considered in the literature.

Hernan etal (2005) for looking at comparing dynamic regimes using artificial censoring.

3.6 Related work

Bryan et al paper fixes the vector L at the beginning which means A never affects L which don't make no sense.

paragraph or two talking about positivity, where it has been studied

3. explain that it has not been thoroughly studied. Cite the examples where it has been. The cole papers but also two empirical applications.

Marginal structural models 1. On MSMs - mainly hernan and Robins work 2. On other work into positivity 3. On Simulating from MSMs

The positivity assumption has received less attention than exchangeability in the literature on MSMs. Two notable exceptions are **?** and **?** who have studied positivity in detail. Both papers aim to give some

Both papers lack a formal simulation study, and neither have looked into detail on the time dependent situation.

The trade-off between positivity and confounding bias is emphasized in Cole2008

Why is practicality important? Cole paper highlights practical advice to practictioners. positivity can be violated in a practical setting because of two few strata, it can be the result of protocols in a clinical setting and it can be seen as a trade-off between exchangeability (and we need more

measured predictors to maintain exchangeability) and positivity where more predictors leads to more likely a zero problem.

Marginal Structural Models

The aim is to be able to simulate the survival function of a desired MSM under the intervention $do(\bar{a})$

Several studies have developed algorithms for simulating data from marginal structural models in the presence of time dependent confounding. An early example is **?** who study estimators of the causal effec of a time dependent treatment on survival outcomes. They compare naive estimators with IPTW and a treatment orthogonalized estimator which is also developed in the paper. This study shares similarities with Havercroft and Didelez (2012)

- stay on treatment after treatment starts
- treatment regime is determined by t* (starting point of treatment because it is a vector of {0, 0, 0, 1, 1, 1}
- they motivate a logistic model for the haxard function, they use a discrete equivalent to the hazrd function (link to citation about farington study.)
- the survival time U is directly linked to the survival outcome -> here it is good to provide more inuition.
- need to understand why it is linked.

  Work that proposes a different method to solve the same problem. Work that uses the same proposed method to solve a different problem. A method that is similar to your method that solves a relatively similar problem. A discussion of a set of related problems that covers your problem domain.

These previous studies have lacked a systematic investigation of positivity violations in a simulation setting. It is unknown, for example, how large an effect a violation of positivity has and how it is affected by the sample siz, threshold etc etc.

1. marginal structural models literature
2. Simulating from marginal structural models
3. positivity
4. connect limited work on simulating from models to positivity
5. Explain why we choose the Havercroft simulation of the others, specifically why it helps us to incorporate violations of positivity in our analysis.

Bryan etal 2010 have a similar focus as Havercroft 2012 in that they develop an algorithm for simulation from a given MSM and they use this algorithm to compare IPTW methods to naive regression methods.

Several studies have considered simulation from margial structural models. The finite-sample properties of marginal structural proportional hazards models has been considered by **?**.

- What is their focus?
- how do they simulate

- what do they find (in terms of MSE, SE, etc.
- how does it differ from HD (2012)

Young etal (2014) also provide a simulation algorithm for simulating longitudinal data from a known Cox MSM.

- What is their focus?

to compare IPW and standard regression based techniques. This is not the subject of this thesis.

Comparing IPW and standard regression based estimates in the absence of model misspecification. This allows for complete isolation of any type of bias. This approach involves simulating data from a standard parametrization of the likelihood and solving for the underlying Cox MSM

- how do they simulate
- what do they find (in terms of MSE, SE, etc.
- how does it differ from HD (2012)

we describe an approach to Cox MSM data generation that allows for a comparison of the bias of IPW estimates versus that of standard regression-based estimates in the complete absence of model misspecification

- could do this section in comparison to Havercroft and didelez, explain how their algorithm works first and then describe the related work by linking differences in their algorithm to earlier or other algorithms.

ALgorithm:

1. generate survival under no treatment from a weibull
2. generate survival times under the ten non-zero treatment levels.
3.

Bryan et al (2010) Cole (2008) for more of a discussion about positivity, with an actual observed data example. While Cole (2008) have looked at positivity in an observational setting, to our knowledge no study has looked at positivity violations within a simulation setting.

Cole and Hernan 2008 have examined the four assumptions underlying IPW using a study on real data from the HAART SWISS study.

1. 2 paper on simulation + Bryan paper - why this simulation method is different from other sim papers
2. related work - Judea pearl, Robins, econometrics, related and broader literature on MSM
3. positivity long discussion in Hernan and cole and the warnings but no simulation study in that paper. They study the positivity but not the effect and there is nothing in the havercroft paper on this either.

-

- Major point is that there are a number of ways of simulating from marginal structural models. But, we need one where we can mess with the positivity. Other methods are not suitable for this.
- G formula simulation
- exposure, confounder feedback loop
- treatment
- outcome variable
- causal effect of the treatment variable on the outcome.
- external intervention
- Explain the do notation, and what presicely is meant in the case

- used to estimate the joint effect of time dependent treatments on survival
- Need to stronly link the time dependent confounding to the MSM, do we choose this class of models because of their relationship with time dep confounding? Yes, marginal strictural models are used with TDC
- The causal graph helps (according to Pearl pp. 40) to bridge statistics into causality
- There is a key part to this, which is that we do not observe confounding, this seems to be what motivates the use of the MSM class of models.
- Counterfactuals need to be addressed here to make it clear this is not the purpose of the thesis.
- Robins (2000) have demonstrated that in the presense of time dependent confounders, standard approaches for asjusting for confounding are biased.
- A covariate that is a risk factor for, or predictor of the event of interest (Y) (from Robins 2000). This defines a time dependent covariate
- And also past exposure determines the level of the covariate.
- Works under a set of assumptions (consistency, exchanchability, positivity and no mispecification of the model used to estimate the weights
-

4 Chapter 2

4.1 Simulating from a static MSM

4.2 Data Structure

We wish to simulate survival data in discrete time $t = 0, \ldots, T$ for n subjects. At baseline $t = 0$ all subjects are assumed to be at risk of failure so that $Y_0 = 0$. For each time period $t = 0, \ldots, T$ a

subject may either be on treatment, $A_t = 1$, or not on treatment, $A_t = 0$. All patients are assumed to be not on treatment before the study begins. Once a patient commences treatment, they remain on treatment in all subsequent periods until failure or the end of follow-up. In each time period L_t is the value of a covariate measured at time t. In the simulated data, L_t behaves in a similar manner to CD4 counts such that a low value of L_t represents a more severe illness and hence a higher probability of both tratemnt and failure in the following period. In addition to L_t, the variable U_t represents subject specific general health at time t. Although we will simulate U_t, in a real world application U_t is an unmeasured confounder which

Each time period is either a check up visit or is between two check up visits. If t is a check-up visit and treatment has not yet commenced, L_t is measured and a decision is made on whether to commence treatment. Between visits, treatment remains unchanged at the value recorded at the previous visit. Similarly, L_t which is only measured when t is a visit, alos remains unchanged.

We represent the history of a random variable with an over bar. For example, the vector representing the treatment history of the variable A is represented by $\bar{A} = [a_0, a_1, \ldots, a_m]$ where $m = T$ if the subject survives until the end of follow-up, or $m < T$ otherwise. Prior to basline both $A = 0$ for all subjects.

- explain what U is and how it relates to the simulation design/algorithm
- Be more specific on Y
- L_t is a measured confounder
- U_t is an unmeasured confounder.

4.3 Simulation Algorithm

4.3.1 Algorithm

Next, we describe the algorithm used to simulate data from our chosen marginal structural model under time dependent confounding. In the following section we discuss in detail how the algorithm works and the salient features for this thesis. The algorithm is taken from Havercroft and Didelez (2012) who generate data on n patients, for k time periods. The outer loop in the following algorithm $i \in 1, \ldots, n$, refers to the patients while the inner loop $t \in 1, \ldots, T$ refers to the subsject specific time periods from baseline to failure or the end of the study. There will be at least one,

and at most T records for each patient.

Algorithm 1: Simulation Algoirthm MSM

Result: Marginal Structural Model Under Time Dependent Confounding

for i *in* $1, \ldots, n$ **do**
- $U_{0,i} \sim U[0,1]$
- $\epsilon_{0,i} \sim N(\mu, \sigma^2)$
- $L_{0,i} \leftarrow F^{-1}_{\Gamma(k,\theta)}(U_{i,0}) + \epsilon_{0,i}$
- $A_{-1,i} \leftarrow 0$
- $A_{0,i} \leftarrow Bern(expit(\theta_0 + \theta_2(L_{0,i} - 500)))$
- **if** $A_{0,i} = 1$ **then**
  - $T^* \leftarrow 0$;
- **end**
- $\lambda_{0,i} \leftarrow expit(\gamma_0 + \gamma_2 A_{0,i})$
- **if** $\lambda_{0,i} \geq U_{0,i}$ **then**
  - $Y_{1,i} \leftarrow 0$
- **else**
  - $Y_{1,i} \leftarrow 1$
- **end**
- **for** $kin1, \ldots, T$ **do**
  - **if** $Y_{t,i} = 0$ **then**
    - $\Delta_{t,i} \sim N(\mu_2, \sigma_2^2)$
    - $U_{t,i} \leftarrow min(1, max(0, U_{t-1,i} + \Delta_{t,i}))$
    - **if** $t \neq 0 \ (mod\ k)$ **then**
      - $L_{t,i} \leftarrow L_{t-1,i}$
      - $A_{t,i} \leftarrow A_{t-1,i}$
    - **else**
      - $\epsilon_{t,i} \sim N(100(U_{t,i} - 2), \sigma^2)$
      - $L_{t,i} \leftarrow max(0, L_{t-1,i} + 150 A_{t-k,i}(1 - A_{t-k-1,i}) + \epsilon_{t,i})$
      - **if** $A_{t-1,i} = 0$ **then**
        - $A_{t,i} \sim Bern(expit(\theta_0 + \theta_1 t + \theta_2(L_{t,i} - 500)))$
      - **else**
        - $A_{t,i} \leftarrow 1$
      - **end**
      - **if** $A_{t,i} = 1 A_{t-k,i} = 0$ **then**
        - $T^* \leftarrow t$
      - **end**
    - **end**
    - $\lambda_{t,i} \leftarrow expit()\gamma_0 + \gamma_1[(1 - A_{t,i})t + A_{t,i}T^*] + \gamma_2 A_{t,i} + \gamma_3 A_{t,i}(t - T^*))$
    - **if** $1 - \prod_{\tau=0}^{t}(1 - \lambda_{\tau,i}) \geq U_{0,i}$ **then**
      - $Y_{t+1,i} = 1$
    - **else**
      - $Y_{t+1,i} = 0$
    - **end**
  - **end**
- **end**

end

Within the inner loop ($t \in 1, \ldots, T$) we see that the data is only updated at time $t \neq 0 \ (mod\ k)$,

where k refers to evenly spaced check-up visits. If t is not a check-up visit the values of A_t and L_t are the same as in $t-1$. When t is a visit A_t and L_t are updated.

- if treatment has been commenced then a subject may feel extra benefit if more time has elapsed since treatment began
- L_t affects A_t and also Y_t
- explain starting values for A and Y are all zero (except L maybe)

In order to operationalize the Algorithm 1 we need to choose parameters for $()$. In their paper Havercroft and Didelez (2012) use values that simulate data with a close resemblance to the Swiss HIV Cohort Study. We postpone disussion of the patameters in Algorithm 1 to section 2.4. We just need to state that we follow their parameters because this is not the focus of this thesis.

4.3.2 Discussion of how algorithm works

The algorithm of Havercroft and Didelez (2012) works by factorizing the joint density of the histories of the four variables in the analysis.

- Important is that the form of the MSM is not specified intil the last stage
- role of $U_{0,i}$
- How does positivity enter the analysis?
- Why this model is important in terms of positivity.

4.4 Constructing IPT weights

Inverse Probability of Treatment weights can be used to adjust for measured confounding and selection bias in marginal structural models. Link back to pseudo population idea in previous section. This method relies on four assumptions consistency, exchangeability, positivity and no mispecification of the model used to estimate the weights Cole and Hernán (2008). Unstabilized weights are defined as:

$$w_{t,i} = \frac{1}{\prod_{\tau=0}^{t} p_\tau(A_{\tau,i} \mid \bar{A}_{\tau-1,i}, \bar{L}_{\tau,i})}$$

Where the denominator is the probability that the subject received the particular treatment history that they were observed to receive up to time t, given their prior observed treatment and covariate histories (Havercroft, Didelez, 2012). The probabilities $p_\tau(A_{\tau,i} \mid \bar{A}_{\tau-1,i}, \bar{L}_{\tau,i})$ may vary greatly between subjects when the covariate history is strongly asscoaited with treatment. In terms of the resulting pseudopopulation, very small values of the unstabilized weights for some subjects would result in a small number of observations dominating the weighted analysis. The result is that the IPTW estimator of the coefficients will have a large variance, and will fail to be normally distributed. This variability can be mitigated by using the following stabilized weights

$$sw_{it} = \frac{\prod_{\tau=0}^{t} p_\tau(A_{\tau,i} \mid \bar{A}_{\tau-1,i})}{\prod_{\tau=0}^{t} p_\tau(A_{\tau,i} \mid \bar{A}_{\tau-1,i}, \bar{L}_{\tau,i})}$$

In the case that there is no confounding the denominator probabiliies in the stabilized weights reduce to $p_\tau(A_{\tau,i} \mid \bar{A}_{\tau-1,i})$ and $sw_{it} = 1$ so that each subject contributes the same weight. In the case of confounding this will not be the case and the stabilized weight will vary around 1.

In practice, we estimate the weights from the data using a pooled logistic model for the numerator and denominator probabilities. The histories of the treatment and covariates are included in the probabilities. In practice Specifically, following Havercroft and Didelez (2012), we estimate the model where the visit is only the visits every check up time. Between check ups both the treatment and covariate remain the same. Other ways of doing this include a spline function over the months to create a smooth function between the visits. Another difference might be to use a coxph function instead of logistic function

$$logit\ p_\tau(A_{\tau,i} \mid \bar{A}_{\tau-1,i}, \bar{L}_{\tau,i}) = \alpha_0 + \alpha_1 k + alpha_2 a_{k-1} + \cdots + alpha_k a_0 +$$

We have several options for estimating these weights. We could use a coxph model, or a logistic model.

4.5 Simulation Set-up

We follow the simulation set-up of Havercroft, Didelez (2012) which is based on parameters that closely match the Swiss HIV Cohort Study (HAART).

4.6 Results

- check the distribution of the weights that come out of the model (see Cole 2008). This would allow us to see weight model mispecifications. Not a problem in the simuation case.
- compare the bias, se, MSE, and 95% confidence interval
- compare all of these in the positivity violation and non-positivty violation case.
- explain to some extent monte-carlo standard error.
- we don't confirm the results of the havercroft of Bryan papers, instead refer readers to these papers to see how IPTW outperforms the naive estimators.

5 Dynamic Case

5.1 The problem of simulating from a MSM under a dynamic strategy

6 Violations of Positivity

The motivation for using the algorith of Havercroft and Didelez (2012) is that we have control over how L affects Y, so we can itroduce positivity using a threshold. In other algorithms there would be a direct link between L and Y, this would be a problem because altering treatment decisions based on L would affect Y directly.
- creating an artificial population in which positivity is violated in specific ways.

6.1 Extended discussion of algorithm linking to positivity

As described in the introduction, one assumption of the model is that there is a non-zero probability of the event occuring at every startum of the covariate.

- When previous covariates like CD4 count are strongly associated with treatment the probabilities in the denominator of the ustabilized weights may vary greatly. Because we are foricing positivity by using a treatment rule when L falls below a threshold and A is then eaual to one, we create a strong association between A and L -> hence the unstabilized weights would vary. (Robins et al 2000 pp. 553)
- present the algorithm again with positivity violations.

7 Application

8 Discussion and Conclusion

8.1 Limitations

References

Yvonne W Cheng, Alan Hubbard, Aaron B Caughey, and Ira B Tager. The association between persistent fetal occiput posterior position and perinatal outcomes: an example of propensity score and covariate distance matching. *American journal of epidemiology*, 171(6):656–663, 2010.

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

WG Havercroft and V Didelez. Simulating from marginal structural models with time-dependent confounding. *Statistics in medicine*, 31(30):4190–4206, 2012.

Lynne C Messer, J Michael Oakes, and Susan Mason. Effects of socioeconomic and racial residential segregation on preterm birth: a cautionary tale of structural confounding. *American journal of epidemiology*, 171(6):664–673, 2010.

Judea Pearl. *Causality*. Cambridge university press, 2009.

James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology, 2000.

Daniel Westreich and Stephen R Cole. Invited commentary: positivity in practice. *American journal of epidemiology*, 171(6):674–677, 2010.