Methodological developments around causal inference and the analysis of high-dimensional data

Lola Etiévant

N^o d'ordre NNT : 2020LYSE1170

THÈSE DE DOCTORAT DE L'UNIVERSITÉ DE LYON
opérée au sein de
l'Université Claude Bernard Lyon 1

École Doctorale N^o 512
InfoMaths

Spécialité du doctorat : Mathématiques

soutenue publiquement le 13 octobre 2020, par :
Lola Étiévant

Développements méthodologiques autour de l'inférence causale et de l'analyse de données en grande dimension

---

Devant le jury composé de :

Julie Josse	Professeure, Inria de Montpellier	Rapporteuse
Stijn Vansteelandt	Professeur, Université de Gand	Rapporteur
Antoine Chambaz	Professeur, Université de Paris	Examinateur
Bianca De Stavola	Professeure, Collège de Londres	Examinatrice
Clément Marteau	Professeur, Université Lyon 1	Examinateur
Franck Picard	Directeur de recherche, CNRS LBBE	Examinateur
Anne-Laure Fougères	Professeure, Université Lyon 1	Directrice de thèse
Vivian Viallon	Maître de conférence, IARC	Directeur de thèse

Et après avis de :

Julie Josse	Professeure, Inria de Montpellier
Stijn Vansteelandt	Professeur, Université de Gand

Institut Camille Jordan

Laboratoire de recherche en mathématiques Lyon/Saint-Étienne

Développements méthodologiques autour de l'inférence causale et de l'analyse de données en grande dimension

Lola Étiévant

Thèse de doctorat

Université Claude Bernard Lyon 1

Université Claude Bernard – LYON 1

Administrateur provisoire de l'Université	M. Frédéric Fleury
Président du Conseil Académique	M. Hamda Ben Hadid
Vice-Président du Conseil d'Administration	M. Didier Revel
Vice-Président du Conseil des Études et de la Vie Universitaire	M. Philippe Chevallier
Vice-Président de la Commission de Recherche	M. Jean-François Mornex
Directeur Général des Services	M. Pierre Rolland

Composantes Santé

Département de Formation et Centre de Recherche et Biologie Humaine	Directrice : Mme Anne-Marie Schott
Faculté d'Odontologie	Doyenne : Mme Dominique Seux
Faculté de Médecine et Maïeutique Lyon Sud - Charles Mérieux	Doyenne : Mme Carole Burillon
Faculté de Médecine Lyon Est	Doyen : M. Gilles Rode
Institut des Sciences et Techniques de la Réadaptation	Directeur : M. Xavier Perrot
Institut des Sciences Pharmaceutiques et Biologiques	Directrice : Mme Christine Vinciguerra

Composantes et Départements de Sciences et Technologie

Département Génie Électrique et des Procédés	Directrice : Mme Rosaria Ferrigno
Département Informatique	Directeur : M. Behzad Shariat
Département Mécanique	Directeur M. Marc Buffat
École Supérieure de Chimie, Physique, Électronique	Directeur : Gérard Pignault
Institut de Science Financière et d'Assurances	Directeur : M. Nicolas Leboisne
Institut National du Professorat et de l'Éducation	Administrateur provisoire : M. Pierre Chareyron
Institut Universitaire de Technologie de Lyon 1	Directeur : M. Christophe Viton
Observatoire de Lyon	Directrice : Mme Isabelle Daniel
Polytech Lyon	Directeur : Emmanuel Perrin
UFR Biosciences	Administratrice provisoire : Mme Kathrin Gieseler
UFR des Sciences et Techniques des Activités Physiques et Sportives	Directeur : M. Yannick Vanpoulle
UFR Faculté des Sciences	Directeur : M. Bruno Andrioletti

Abstract

Cancer epidemiology is concerned with the identification of causes of cancer, including the biological mechanisms possibly involved in cancer development, based on observational data. The tools recently introduced in the causal inference literature offer a formal framework to address such causal questions. In particular counterfactual variables can be used to define causal effects of interest, and different sets of conditions have been shown to be sufficient to guarantee that a given causal effect can be estimated in practice. However, the practical application of causal inference in cancer epidemiology still faces a number of challenges; the objective of this thesis is to explore some of them.

First, concerns have been raised in the literature regarding the relevance of causal effects estimated from observational studies for certain exposures, for instance obesity, for which there is no possible "direct" intervention, and only interventions on some of its causes, such as diet or physical activity, can be implemented in practice. We show how the effect of an hypothetical intervention on the exposure of interest, when impossible to apply in practice, relates to the effects of interventions on its causes, depending on the structure of the causal model.

Then, even if many causal models of interest in epidemiology involve time-varying variables, these variables are most often observed at one single point in time only. Then, practitioners tend to overlook the time-varying nature of the variables, and to work under over-simplified causal models. We investigate conditions ensuring that estimates derived under over-simplified longitudinal causal models relate to causal quantities of interest under the true longitudinal causal model. Our results confirm that these conditions are very stringent and that estimates derived under over-simplified longitudinal causal models generally have to be interpreted with great caution.

Motivated by a project on high-dimensional mediation analysis, we also study latent variable models for dimension-reduction. We notice a severe limitation in several models proposed in the literature, including the probabilistic formulation of partial least squares proposed by el Bouhaddani et al. (2018). We precisely describe the limitation under this particular model, and illustrate it through simulated examples. We also propose a simple extension which corrects the defect of the initial model. Overall, our results suggest that caution is needed when developing and applying latent variable models for dimension-reduction, as they may turn out to be too simplistic when imposing too strong constraints on the model parameters.

Finally, with the same motivating example in mind, we study the calibration of the tuning parameter in penalized regression models. We focus on a popular extension of the lasso, the adaptive lasso, which uses a weighted L_1-norm in the penalty term, with weights derived from initial estimates of the parameter vector. We empirically show that the standard K-fold cross-validation, although very popular, is not suitable to calibrate the tuning parameter in the adaptive lasso. A simple alternative cross-validation scheme is proposed, which is shown to outperform the standard K-fold cross-validation on simulated examples.

Résumé

L'identification des causes du cancer, mais aussi des mécanismes biologiques pouvant intervenir dans son développement, à partir de données observationnelles, est l'une des problématiques principales en épidémiologie du cancer. Les outils introduits récemment en inférence causale offrent un cadre formel pour répondre à de telles questions. En particulier, les variables contrefactuelles permettent de définir les effets causaux d'intérêts, et diverses conditions permettent de garantir qu'un effet causal donné soit estimable en pratique. Cependant, leur mise en application en épidemiologie du cancer présente un certain nombre d'enjeux ; l'objectif de cette thèse est d'en explorer quelques uns.

Tout d'abord, des réserves ont été émises concernant la pertinence des effets causaux estimés à partir de données observationnelles pour des expositions telles que l'obésité, pour laquelle il n'existe pas d'intervention «directe», mais seulement des interventions sur certaines de ses causes, comme l'activité physique ou l'alimentation. À cet effet, nous étudions comment l'effet d'une intervention hypothétique sur l'exposition d'intérêt est lié aux effets des interventions sur certaines de ses causes.

Ensuite, même si la plupart des modèles causaux d'intérêt en épidémiologie font intervenir des variables qui varient au cours du temps, ces dernières ne sont bien souvent observées qu'à un unique temps donné. De fait, il est assez usuel de travailler sous un modèle causal simplifié, qui néglige le caractère longitudinal de ces variables. Nous déterminons des conditions qui assurent que les quantités obtenues en travaillant sous de tels modèles soient liées à celles d'intérêt sous le vrai modèle longitudinal. Ces conditions, très restrictives, confirment ainsi que les quantités obtenues en travaillant sous des modèles causaux longitudinaux simplifiés doivent généralement être interprétées avec prudence.

Motivé.e.s par un projet sur les analyses en médiation en grande dimension, nous nous sommes intéressé.e.s à l'utilisation des modèles à variables latentes pour la réduction de dimension. Nous avons identifié un défaut dans plusieurs modèles proposés dans la littérature, notamment dans la formulation probabiliste des moindres carrés partiels proposée par el Bouhaddani et al. (2018). Nous décrivons en détail le défaut sous leur modèle, et l'illustrons au moyen de simulations. Nos résultats suggèrent que les modèles à variables latentes doivent être développés avec précaution pour faire de la réduction de dimension, puisqu'ils peuvent en fait être trop simples lorsque les contraintes imposées sur les paramètres sont trop fortes.

Enfin, toujours motivé.e.s par le même projet, nous nous intéressons à la sélection du paramètre de régularisation dans les modèles de régression pénalisés. Plus précisément, nous considérons le lasso adaptatif, une extension du lasso qui utilise une version pondérée de la norme L_1 dans le terme de pénalité, où les poids sont obtenus à partir d'une estimation initiale du vecteur de paramètres. Nous montrons de manière empirique que la validation croisée «K-fold», bien que couramment employée, n'est pas adaptée à la calibration du paramètre de régularisation pour le lasso adaptatif. Une procédure alternative est proposée, et nous montrons sur des simulations qu'elle présente de meilleures performances que la validation croisée «K-fold».

Résumé substantiel

Reposant largement sur des études observationnelles, l'épidémiologie vise à étudier l'effet causal de différents facteurs sur la survenue de pathologies, en particulier les cancers, qui sont l'une des principales causes de mortalité dans le monde. Comme la plupart des maladies chroniques, les cancers ont des causes multiples, qui peuvent interagir. Ainsi, les épidémiologistes portent un intérêt grandissant aux analyses en médiation, qui permettent une description fine des mécanismes d'action de causes de cancers. Grâce à elles, on peut par exemple espérer décrire les mécanismes expliquant le rôle de l'obésité ou du style de vie (activité physique, consommation d'alcool, etc.) sur le développement de cancers (foie, sein, etc.), en considérant des métabolites comme médiateurs potentiels. Cependant d'un point de vue formel, l'étude statistique des effets causaux se heurte à un problème fondamental : les outils classiques en statistique ne permettent que d'étudier l'association entre deux variables, or l'existence d'une association entre, par exemple, une exposition d'intérêt X et l'indicatrice Y de survenue d'un cancer n'implique évidemment pas que X soit une cause de Y. Cette association pourrait aussi s'expliquer par le fait que Y soit une cause de X, ou par l'existence d'une cause commune à X et Y, etc. Notamment dans l'exemple où X est l'obésité et Y l'occurence du cancer, ces causes communes, ou facteurs de confusion, peuvent inclure l'activité physique et le régime alimentaire : la relation entre X et Y peut être confondue, et dans ce cas, ne peut être interprétée de manière causale. Afin de formaliser le concept de causalité, une littérature conséquente a emergé récemment. L'inférence causale repose notamment sur la notion de variables contrefactuelles (Rubin, 1974), qui permettent de définir les effets causaux d'intérêts, que ce soit l'effet causal total d'une exposition (Rosenbaum and Rubin, 1983), ou encore sa décomposition en la somme d'un effet naturel direct et d'un (ou plusieurs) effet(s) naturel(s) indirect(s), médié(s) par de potentiels médiateurs (Pearl, 2001, Robins and Greenland, 1992). Diverses conditions, garantissant qu'un effet causal donné soit estimable à partir de données observationnelles, ont aussi été proposées dans la littérature. Cependant l'application, notamment en épidémiologie du cancer, des outils introduits en inférence causale se heurte en pratique à un certain nombre d'enjeux. L'objectif de cette thèse est d'en étudier certains.

Le premier projet porte sur l'interprétation des effets causaux estimés à partir de données observationnelles pour des expositions pour lesquelles il n'existe pas d'intervention «directe», mais seulement des interventions sur certaines de leurs causes. C'est notamment le cas lorsque l'exposition d'intérêt est l'obésité : dans ce cas, il n'est pas possible d'intervenir directement pour modifier le niveau d'exposition d'un individu, et seules des interventions sur ses causes sont possibles, comme le régime alimentaire ou l'activité physique par exemple. Ainsi, nous étudions comment l'effet d'une intervention hypothétique sur l'exposition d'intérêt, lorsque celle-ci n'est pas réalisable en pratique, est lié aux effets des interventions sur certaines de ses causes. Pour cela, nous nous appuyons sur la structure du modèle causal, et à titre d'exemple, nous supposons plus précisément que l'exposition d'intérêt est l'obésité à l'âge de 20 ans. En particulier, puisque la plupart des causes modifiables de l'obésité sont des facteurs de confusions pour sa relation

avec le cancer, l'effet de l'obésité, estimé à partir de données observationnelles, diffère très probablement des effets d'une intervention sur ses causes. Sous certaines hypothèses sur le modèle causal, il peut par ailleurs être vu comme un effet indirect d'interventions particulières sur ces causes.

Le second projet porte sur l'inférence causale sous des modèles causaux longitudinaux simplifiés. En effet, la plupart des modèles causaux d'intérêt en épidémiologie font intervenir des variables d'exposition (activité physique, consommation d'alcool etc.), ainsi que des médiateurs et facteurs de confusion potentiels, qui varient au cours du temps. Cependant, alors que ces modèles causaux sont ainsi longitudinaux, les données disponibles dans les cohortes pour ces différents facteurs ne concernent généralement que leur niveau à l'inclusion dans l'étude. De fait, il est assez usuel de travailler sous un modèle causal simplifié, où le caractère longitudinal des variables est négligé. Nous étudions alors si, et le cas échéant comment, les quantités obtenues en travaillant sous de tels modèles simplifiés sont liées à celles d'intérêt sous le vrai modèle longitudinal. Plus précisément, nous nous concentrons sur deux cas de figure, lorsque les données disponibles correspondent (i) à des niveaux instantanés mesurés à l'inclusion dans l'étude, ou (ii) à des mesures résumées de l'historique d'exposition jusqu'à l'inclusion. Ainsi, nous déterminons des conditions qui assurent que les quantités obtenues en travaillant sous des modèles causaux longitudinaux simplifiés puissent s'exprimer comme un effet causal longitudinal d'intérêt, ou comme une moyenne pondérée de tels effets causaux. Cependant, puisque l'interprétabilité de ces moyennes pondérées n'est pas toujours évidente, et puisque les conditions déterminées sont très restrictives, nos résultats confirment que les quantités obtenues en travaillant sous des modèles causaux longitudinaux simplifiés doivent généralement être interprétées avec prudence.

Les deux projets suivants ont été motivés par l'analyse en médiation en grande dimension. En effet, de nombreuses études en épidémiologie du cancer visent actuellement à identifier les métabolites, notamment, qui pourraient expliquer l'effet carcinogène de l'obésité, ou d'autres facteurs liés au mode de vie, sur plusieurs types de cancer. Les jeux de données à disposition sont ainsi de relativement grande dimension, et peuvent contenir des métabolites fortement corrélés les uns avec les autres; c'est notamment le cas dans l'étude EPIC sur le cancer de l'endomètre, où environ 150 métabolites ont été mesurés. À cet effet, nous nous sommes en particulier intéressé.e.s à l'utilisation des modèles à variables latentes pour la réduction de dimension. Des formulations probabilistes de certaines techniques de réduction de dimension, comme l'analyse en composantes principales ou les moindres carrés partiels (PLS), ont été proposées dans la littérature, et très récemment cette idée a été étendue aux modèles de médiation. En développant un modèle à variables latentes adapté aux analyses en médiation de grande dimension où l'ensemble d'exposition est lui aussi multivarié, nous avons cependant identifié un défaut, qui est en fait également présent dans d'autres modèles probabilistes. C'est en particulier le cas dans le modèle de PLS probabiliste (PPLS) proposé par el Bouhaddani et al. (2018). Nous décrivons en détail le défaut sous leur modèle, et montrons que leurs contraintes sur les

paramètres sont telles que le modèle définit un ensemble de lois de probabilité très particulières, où les composantes de covariance maximale sont nécessairement aussi de variances maximales, respectivement. Nous illustrons ce défaut au moyen de simulations, et proposons aussi une extension du modèle, pour obtenir une formulation plus "générale" et qui n'est pas limitée à ces seules lois. Nos résultats suggèrent que les modèles à variables latentes doivent être développés avec précaution pour faire de la réduction de dimension, puisqu'ils peuvent en fait perdre leur intérêt apparent lorsque les contraintes imposées sur les paramètres sont trop fortes.

Enfin dans un dernier projet, nous nous intéressons au problème de la calibration du paramètre de régularisation dans les modèles de régression pénalisés, qui sont couramment employés pour l'analyse de données de grande dimension. Notamment dans le modèle que nous avions envisagé pour l'analyse en médiation de grande dimension, nous avions considéré un algorithme d'estimation pénalisé par la norme L_1 afin d'encourager certains paramètres à être creux. Ici, nous nous intéressons plus particulièrement au lasso adaptatif, une extension du lasso qui remplace la norme L_1 dans le terme de pénalité par une version pondérée, où les poids sont obtenus à partir d'une estimation initiale du vecteur de paramètres. La méthode couramment utilisée dans ce cas pour sélectionner la valeur du paramètre de régularisation est la validation croisée, dite validation croisée «K-fold». Nous montrons de manière empirique que dans le cadre des modèles de régression linéaire, la validation croisée «K-fold» n'est pas adaptée à la calibration du paramètre de régularisation pour le lasso adaptatif. Une procédure de calibration alternative est proposée, et nous montrons finalement sur une étude de simulations qu'elle présente de meilleures performances que la validation croisée «K-fold».

Remerciements

Je voudrais tout d'abord remercier Vivian Viallon, pour m'avoir encadrée, accompagnée et encouragée pendant ces trois années de thèse. Je suis très chanceuse d'avoir pu travailler avec toi, d'avoir pu bénéficier de tes idées, de ta motivation et de ta bienveillance ; tu m'as permis de progresser et de me dépasser, mais aussi de ne pas (trop) me laisser submerger par le stress. Je voudrais ensuite remercier Anne-Laure Fougères, pour ses conseils, ses encouragements et bien sûr sa grande bienveillance ; j'ai aussi beaucoup aimé partager mes enseignements avec toi, malgré les circonstances pas toujours évidentes cette dernière année. Je ressors enrichie scientifiquement et humainement de cette belle expérience avec vous deux.

Je voudrais ensuite remercier Antoine Chambaz, Bianca De Stavola, Clément Marteau et Franck Picard pour avoir accepté de faire partie de mon jury de thèse. Je remercie aussi Julie Josse et Stijn Vansteelandt, qui ont de plus accepté et pris le temps de rapporter ma thèse, et m'ont permis de bénéficier de leurs précieuses et constructives remarques.

Je remercie les membres de mon comité de suivi de thèse, Anne Gégout-Petit et Thomas Lepoutre, pour leur bienveillance et leurs encouragements. Un grand merci à Thibault Espinasse, Edouard Ollier et Franck Picard, qui m'ont prodigué de précieux conseils et accordé de leur temps, en particulier Edouard qui n'a pas hésité à venir plusieurs fois depuis St Etienne. Merci également à Frédéric Lagoutière.

Je souhaite ensuite remercier les personnes que j'ai eues la chance de rencontrer lors de mon stage et de mes visites au Centre International de Recherche sur le Cancer, en particulier Flavie et Sabine pour leur bonne humeur, leur symphatie et leurs nombreux conseils, mais aussi Ghazaleh et Hannah qui m'ont rassurée et aidée sur ma préparation à l'après thèse.

Du côté de l'Institut Camille Jordan, je voudrais remercier Simon Masnou pour son écoute et les échanges que nous avons eus, notamment autour de notre projet de charte. Merci aussi à Luca Zamboni pour avoir écouté les propositions diverses des représentant.e.s des doctorantes et doctorants. Merci à Christine Le Sueur et Lydia Barlerin pour leur aide, aussi bien administrative que culinaire avec les divers gouters qui ont été organisés. Merci à Benoit Fabrèges, Laurent Azema, Roland Denis et Vincent Farget pour leur aide «technique». Je remercie enfin chaleureusement les doctorantes et doctorants de l'Institut Camille Jordan pour leur convivialité. Merci en particulier à mes collègues du bureau 100, João, Jorge, Marina, Maxime, Maxime, Mélanie, Pan, Rémy, Sam, Simon, Vincent, mais aussi Gwladys et Marion. Merci à Sally, Théo et Vincent pour la co-organisation du séminaire des doctorantes et doctorants l'année dernière. Merci à mes collègues représentants de l'École Doctorale Info-Maths, Hugo tout d'abord, puis Clément. Merci aux membres du bureau des doctorantes et doctorants, Caterina, Marina et Octave, avec qui nous avons passé de nombreuses heures à échanger et travailler !

Je remercie bien sûr ma famille, mes parents, Marie-Christine et Patrick, ma soeur Sibyle, mais aussi Cécile et Jalel, mes grand parents Claudine et Maurice, mes oncles et tantes et mes cousins et cousines. Je remercie mes ami.e.s, qui m'ont permis de me changer les idées mais

aussi d'échanger sur cette expérience particulière qu'est la thèse, en particulier Lucie, Mathilda, Maria et Soihaila. Je remercie enfin Sémi, soutien sans faille avant et pendant ces trois années, qui m'a supportée et encouragée.

Merci à celles et ceux qui de près ou de loin m'ont aidée à avancer lors de ces trois années.

Contents

Chapter 1

Introduction

1.1 Cancer epidemiology and causal inference

1.1.1 Context

Cancer is the second leading cause of death worldwide, accounting for almost 10 millions deaths in 2018 (Wild et al., 2020). In the United States, it is estimated that about 40% of the men and 38.7% of the women will develop cancer over the course of their lifetime (Howlader et al., 2020). As most other chronic conditions, cancer is a multifactorial disease: causes of most site-specific cancers (breast, prostate, lung, colon, head and neck, etc.) are numerous, and generally combine both genetic and "environmental" (i.e., non-genetic) factors. In this context, one of the main objectives of cancer epidemiology is the identification of the causes of cancer, especially the modifiable ones, in order to devise efficient risk-reduction policies. For example, there is increasing evidence of a causal effect of various lifestyle factors, including tobacco smoking, alcohol and obesity, on the risk of several site-specific cancers (Agudo et al., 2012, Bagnardi et al., 2015, Lauby-Secretan et al., 2016). Another recent line of research in cancer epidemiology aims at investigating the biological mechanisms underlying the carcinogenic effect of these lifestyle factors (Khandekar et al., 2011, Renehan et al., 2015). Several studies have recently focused on the carcinogenic effect of obesity, suggesting that it could be partly *mediated* by chronic inflammation and insulin resistance: more precisely, obese people are more likely to suffer from chronic inflammation and insulin resistance, which, in turns, may increase cancer risk (Dashti et al., 2019, 2020). Despite its interest by itself, the description of these biological mechanisms could also lead to prevention interventions, in particular when an intervention on the biological mechanisms (e.g. chronic inflammation) is easier than an intervention on the primary exposure (obesity). Rather than focusing on a few candidates mechanisms, cancer epidemiologists can now adopt a more agnostic approach, and explore biological mechanisms possibly underlying some exposure-cancer relationships by studying different -omics data, which are now available in large epidemiological studies. In par-

ticular, metabolomics data are available in a number of studies, including the European Prospective Investigation into Cancer and nutrition (EPIC) study. Metabolomics data consist of measures on a broad panel of molecules (metabolites). It is supposed to provide a good description of the complete "metabolome" of an individual at a given point in time, and can be seen as "a readout of the integrated response of cellular processes to genetic and environmental factors" (Krumsiek et al., 2011). On the one hand, epidemiological studies using metabolomics data have identified metabolites related to several cancers (His et al., 2019). On the other hand, a number of metabolites have been shown to be associated with several lifestyle exposures such as physical activity, alcohol consumption and smoking status, but also obesity (Cirulli et al., 2019, Du et al., 2020, Harada et al., 2016, Langenau et al., 2019). Then, a few epidemiological studies have lately focused on the evaluation of the mediating role of metabolites available in metabolomics data in the relationship between different lifestyle factors and cancer sites (Assi et al., 2015a,b, Petimar et al., 2018).

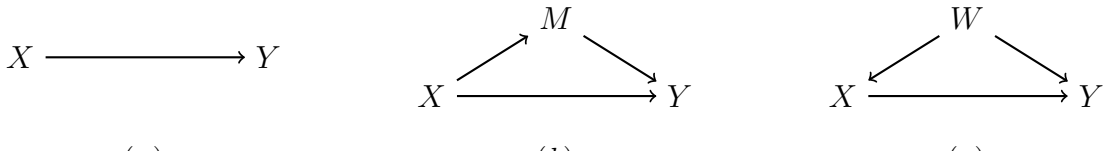
1.1.2 From statistical association to causal effect

To address these various questions, cancer epidemiology relies on the statistical analysis of data which can originate from experimental, or more often, observational studies. Randomized clinical trials are one particular variety of experimental studies, where the level of the exposure or treatment of interest of each subject is randomly assigned. Consider for simplicity a binary exposure; thanks to the randomization, the only difference between the two groups (exposed and non-exposed) is, in principle, their exposure level, so that if a difference exists between the cancer risks in the exposed and unexposed groups, it can generally be interpreted causally (Ahrens and Pigeot, 2005). However, most epidemiological results are derived from observational data, for which association does not imply causation. Indeed, an observed association between two variables X and Y, for example obesity and cancer occurrence, does not imply that X is a cause of Y. This association could be due to the fact that Y is a cause of X (reverse causation), or due to the existence of a common cause of X and Y (confounding). In the example where X and Y represent obesity and cancer occurrence, respectively, these common causes, or confounders, may include diet and physical activity: the $X - Y$ relationship may be confounded, at least partly, and if so, cannot be interpreted causally. Other more subtle mechanisms (e.g. selection bias) can also result in spurious (non-causal) association between two variables. On the one hand, the identification of so-called risk factors, X, associated with cancer occurrence, can be sufficient to address some epidemiological questions, such as the prediction of cancer risk. But, on the other hand, to devise efficient risk-reduction policies, it is necessary to determine whether or not these factors are genuine causes of cancer, to make sure that by improving the exposure levels in the population (e.g. reducing adiposity), cancer risk would be reduced.

Lately, a significant literature, devised jointly by the statistics, mathematics and informatics communities, has emerged to formalize the concept of causality. The standard approach for causal inference from observational data involves counterfactual variables (Rubin, 1974), as well as Structural Causal Models (SCMs) (Pearl, 1995, 2000), which are based on the probabilistic graphical model theory, and deeply rely on Directed Acyclic graphs (DAGs) and *d*-separation (Lauritzen, 1996, Verma and Pearl, 1988).

In particular, counterfactual variables allow the formal definition of causal quantities of interest. For example, consider the case where X, the binary exposure under study, is the obesity status (obese/lean), and Y, the outcome of interest, is cancer occurrence. To assess the causal effect of obesity on cancer occurrence at the population level (on average), we would like to compare two cancer risks: (i) the risk of cancer in the population, had all the individuals of that population been obese, and (ii) the risk of cancer in the population, had all the individuals of that population been lean. Of course, these two populations are different from the actual population: they are counterfactual. And we will never be in the position to have access to data from these two populations. But, we can still formally define the variables that we could have observed if we have had access to data from these counterfactual populations. Let $do(X = x)$ denote an hypothetical intervention such that the exposure X is forced to take value x. Then, denote $Y^{X=x}$ the outcome variable that would have been observed in the counterfactual world (or population) following the intervention $do(X = x)$; just as the counterfactual population, this variable is not observed in the real world: it is a counterfactual variable. The average causal effect of X on Y can still be formally defined by comparing the distributions of $Y^{X=1}$ and $Y^{X=0}$. In particular, on the additive scale, the average, or total, causal effect of X on Y can be defined as $\mathbb{E}(Y^{X=1} - Y^{X=0})$. Yet, this quantity is defined in terms of $Y^{X=1}$ and $Y^{X=0}$, which are unobserved. A natural question is whether the causal effect can be expressed in terms of the distribution of observable variables, X, Y and potentially additional variables. In particular, it is noteworthy that $\mathbb{E}(Y^{X=x})$ usually differs from $\mathbb{E}(Y \mid X = x)$; see Appendix A for more details. When a causal effect can be expressed as a function of the distribution of observed variables, it is said to be "identifiable". Counterfactual variables further allow to derive sets of sufficient conditions, namely consistency, (conditional) ignorability and (conditional) positivity, under which total causal effects can be identified.

However, these conditions are not (fully) testable, and causal inference then usually relies on some prior knowledge, such as the structure of the causal system of interest. For example, Pearl's SCMs notably combine graphical causal models and sets of structural equations, to describe and specify our assumptions or knowledge of the possible relationships among the variables involved in the causal system under consideration. A graphical causal model is typically a DAG, as in Figure 1.1 (a), (b) or (c), where each node corresponds to a variable of the causal system, and where the possible existence of a causal relationship between two variables is translated into a directed edge between the



(a) (b) (c)

Figure 1.1: (a) Example of a graphical causal model where the exposure X is a potential cause of the outcome Y. (b) Example of a graphical causal model where the effect of the exposure X on outcome the Y is possibly mediated by M. (c) Example of a graphical causal model where the effect of the exposure X on the outcome Y is possibly confounded by W.

corresponding nodes. For example, in Figure 1.1 (a), the presence of a directed edge from X to Y means that X may be cause of Y, and not the other way around; see Appendix A for further details on causal models and SCMs.

The question of the identifiability of causal effects can be systematically addressed under the frameworks of SCMs. For example, a fundamental result is that, if all of the so-called endogeneous variables are observed, then the total causal effect of any set of variables on any other disjoint set of variables is always identifiable (Pearl, 2000). On the other hand, when some endogeneous variables in the model are not observed, the causal quantity of interest is not necessarily identifiable, but several graphical criteria, such as the front-door and back-door criteria (Pearl, 1993, 1995), have been shown to be sufficient, in the sense that if they are satisfied, then the identifiability of the causal effect is guaranteed. In particular, the back-door criterion allows the identification of a set of variables W such that the conditional ignorability condition $\big(Y^{X=x} \perp\!\!\!\perp X \mid W\big)$ holds, and $\mathbb{E}(Y^{X=x})$ can be expressed as $\sum_w \mathbb{E}\big(Y \mid W = w, X = x\big) \times \mathbb{P}\big(W = w\big)$, with the sum over all possible values of W (which is assumed to be categorical here, for simplicity). A generalization of these criteria has been proposed through several necessary and sufficient graphical conditions (Shpitser and Pearl, 2006, Tian and Pearl, 2002); we refer to Appendix A for a brief summary of some of these identifiability criteria.

The tools briefly recalled above can be useful to assess, e.g., the causal effect of obesity on cancer. They have further been extended to cover mediation analyses, which can be used to investigate whether the causal relationship between an exposure X and outcome Y can be partly explained, or mediated, by another variable. Consider for example the graphical causal model given in Figure 1.1 (b), where X affects Y both "directly" and "indirectly", through M. If this is the case, we say that M is a mediator in the $X - Y$ relationship. Formally, the objective of mediation analysis is to quantify the portion of the total causal effect of X on Y that is mediated by M, the indirect effect, and the portion that is not mediated by M, the direct effect. In particular, it was shown that the total causal effect can be decomposed into the sum of the so-called natural direct and indirect effects (Pearl, 2001, Robins and Greenland, 1992), which are again both formally defined from counterfactuals variables. Sufficient conditions have been proposed in the

literature, which ensure the identifiability of the natural direct and indirect effects ; see Pearl (2001), VanderWeele (2015) and the Appendix A for more details.

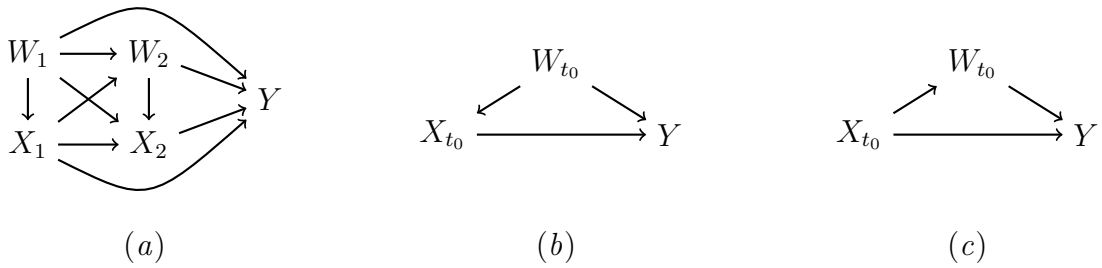
To recap, a formal framework, based on a number of tools recently introduced in the causal inference literature, is now available to address causal queries of interest in cancer epidemiology. In particular, these tools provide a precise definition of various concepts, such a confounding or selection bias, that have "always" been described in epidemiology and biostatistics, but with a rather intuitive, and less precise, definition. Yet, the practical application of causal inference and mediation analyses in cancer epidemiology still faces several challenges, and the objective of this thesis was to explore some of them; they are summarized in Section 1.2.

1.2 Thesis objectives and main contributions

1.2.1 The practical interventions the *do*-operator refers to in causal inference; illustration on the example of obesity and cancer

Consider the case where X, the binary exposure under study, is the obesity status, say at the age of 20, and Y, the outcome of interest, is cancer occurrence. The causal effect of X on Y can be defined as $E(Y^{X=1} - Y^{X=0})$, where $Y^{X=1}$ and $Y^{X=0}$ are the outcome variables that would have been observed in the counterfactual worlds following the interventions $do(X = 1)$ and $do(X = 0)$, respectively. However, interventions $do(X = 1)$ and $do(X = 0)$ are not unique, and hence not well-defined, as they could correspond to any dynamic interventions (Daniel et al., 2012, Hernán and Robins, 2020) ensuring that individuals stay lean and get obese by the age of 20, respectively. In particular, in order to prevent obesity by the age of 20, individuals could be asked to do 45 minutes of physical exercise a day, or 72 minutes of physical exercise a day, or they could also be asked to adhere to a healthy diet, etc.

This situation, where several practical interventions on the causes of the obesity could lead to a same obesity level, falls under the general case of a treatment with multiple versions (Hernán and VanderWeele, 2011, Petersen, 2011, VanderWeele and Hernán, 2013), and then violates the "no-multiple-versions-of-treatment assumption", which is part of the "Stable Unit Treatment Value Assumption" (Rubin, 1980, VanderWeele and Hernán, 2013). Concerns have been raised in the literature regarding the relevance of causal effects estimated from observational studies in such cases, but most arguments have been based on situations where "treatment precedes versions of that treatment", while situations where "versions precede treatment" were only quickly mentioned, if ever (Hernán and VanderWeele, 2011, Petersen, 2011, VanderWeele and Hernán, 2013). In Chapter 2 we investigate how the effect of an hypothetical intervention on the exposure of interest, when impossible to apply in practice, relates to the effects of interventions on its causes.



(a) (b) (c)

Figure 1.2: (a) Example of a longitudinal causal model, with time-varying exposure $(X_t)_{t\in[\![1;t_0]\!]}$ and time-varying confounder $(W_t)_{t\in[\![1;t_0]\!]}$ affected by the exposure. The causal diagram is given for $t_0 = 2$. (b) Over-simplified causal model associated with the longitudinal model given in Figure 1.2 (a). (c) Another possibility of over-simplified causal model associated with the longitudinal model given in Figure 1.2 (a).

Following the recommendations of Petersen (2011), the investigation relies on the structural causal model framework. With the example of X standing for obesity at the age of 20 in mind, we explore different scenarios, including the situation where some causes of X are modifiable, while others are not (e.g. genetic determinants of obesity).

1.2.2 Causal inference under over-simplified longitudinal models

Consider once again the study of the causal relationship between obesity and cancer. Because insufficient physical activity is likely a cause of both obesity and cancer, physical activity is usually considered as a confounder in the obesity-cancer relationship. But on the other hand, obesity is also likely to decrease physical activity, and then physical activity could be considered as mediator in the relationship between cancer and obesity. Graphical causal models given in Figure 1.1 (b) or (c) are actually too simplistic to properly describe the relationship between obesity and physical activity. Indeed, the true causal model involves time-varying variables: it is a longitudinal model, as the one depicted in Figure 1.2 (a) for $t_0 = 2$, where $(X_t)_{t\in[\![1;t_0]\!]}$ could stand for the obesity status at different ages, while $(W_t)_{t\in[\![1;t_0]\!]}$ could stand for physical activity at different ages. Notably, X_t, the exposure variable at time t, affects the confounding variable $W_{t'}$ at any time $t' > t$, as well as its own future value $X_{t'}$. Then $(W_t)_t$ is a so-called time-varying confounder affected by the exposure, and roughly speaking, it is both a confounder and a mediator in the $(X_t)_t - Y$ relationship.

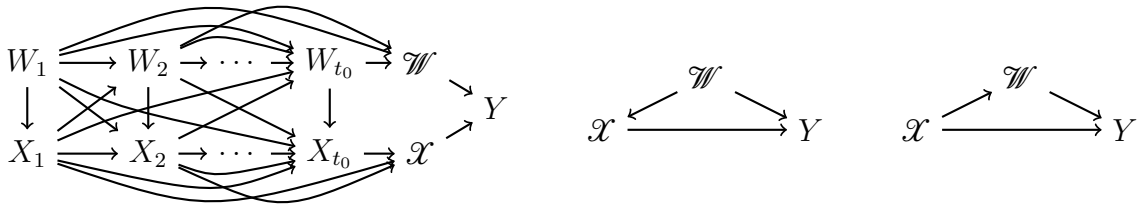
The tools of causal inference for time-fixed variables have been extended to such longitudinal settings (Pearl and Robins, 1995, Robins, 1986, VanderWeele, 2015). In particular, in the case where the causal effect of interest is that of a binary exposure varying over some the discrete time interval $[\![1; t_0]\!]$, on the outcome Y measured at some later point in time $T > t_0 > 1$, let $\bar{X}_{t_0} = (X_1, \ldots, X_{t_0})$ denote the exposure profile until time t_0. Then the causal effect can be formally defined by $\mathbb{E}\left(Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} - Y^{\bar{X}_{t_0}=\bar{x}^*_{t_0}}\right)$, for two given profiles $\bar{x}_{t_0} = \left(x_1, \ldots x_{t_0}\right)$ and $\bar{x}^*_{t_0} = \left(x^*_1, \ldots x^*_{t_0}\right)$, with $Y^{\bar{X}_{t_0}=\bar{x}_{t_0}}$ the outcome variable that would

have been observed in the counterfactual world following $do(\bar{X}_{t_0} = \bar{x}_{t_0})$. This causal effect is the total causal effect of the exposure until time t_0 on the outcome Y, and for example with $\bar{x}_{t_0} = (1, \ldots, 1)$ and $\bar{x}^*_{t_0} = (0, \ldots, 0)$, it corresponds to the difference between the risks that would have been observed in the two populations where all individuals would have been "always obese" and "never obese", respectively. Under certain conditions, including the "sequential" ignorability condition, which can be seen an extension of the conditional ignorability condition to longitudinal settings (Daniel et al., 2012, Hernán and Robins, 2020, Robins, 1986), identifiability of longitudinal total causal effects is guaranteed. For instance under the causal model of Figure 1.2 (a), $\mathbb{E}\left(Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} - Y^{\bar{X}_{t_0}=\bar{x}^*_{t_0}}\right)$, for any given profiles $\bar{x}_{t_0}$ and $\bar{x}^*_{t_0}$, can be expressed in terms of Y, $\bar{X}_{t_0}$ and $\bar{W}_{t_0} := (W_1, \ldots W_{t_0})$ only. We refer to Appendix A for more details on longitudinal causal inference.

Repeated measurements for time-varying exposures, and possibly mediators and confounders, are then usually required to perform valid causal inference under longitudinal causal models. Yet, they are rarely available in large observational studies, as for instance in the EPIC study and the UK Biobank (Riboli et al., 2002, Sudlow et al., 2015), where most variables, in particular those measured from blood samples, are usually collected once only, at recruitment typically. Some of the general results on the identifiability of causal effects in the presence of unobserved variables in the causal model mentioned above (Shpitser and Pearl, 2006, Tian and Pearl, 2002) could be used to study the identifiability of the causal effect of interest when ignoring the time-varying nature of exposures or, equivalently, when past levels of exposures are unobserved. However, they are generally not considered in practice, and practitioners tend to directly work under over-simplified causal models, where the time-varying nature of exposures, mediators and confounders are simply ignored (Bradbury et al., 2019, Chan et al., 2011, Dossus et al., 2013, Petimar et al., 2018, Schairer et al., 2016).

Even if issues arising when working under over-simplified longitudinal causal models have already been described in the statistical literature (Aalen et al., 2016, Maxwell and Cole, 2007, Maxwell et al., 2011), little is known about the relationship between estimates derived under over-simplified longitudinal causal models and causal quantities of interest under the true longitudinal causal model. In Chapter 3 we investigate conditions ensuring that the quantity estimated in practice when working under over-simplified longitudinal causal models expresses as a particular weighted average of the longitudinal causal effects of interest. More precisely, we consider two different situations regarding the available data for the "exposures", which include the exposure of interest but also possibly some mediators and confounders.

The first and most common situation is when available data for the exposures correspond to their "instantaneous" levels at the time t_0 of recruitment in the study. For example, under the causal model depicted in Figure 1.2 for $t_0 = 2$, only data on Y, X_{t_0} and W_{t_0} would be available, while data on $\bar{X}_{t_0-1}$ and $\bar{W}_{t_0-1}$ would not. Then, over-



(a) (b) (c)

Figure 1.3: (a) Example of longitudinal causal model, with time-varying exposure $(X_t)_{t\in[\![1;t_0]\!]}$ and time-varying confounder $(W_t)_{t\in[\![1;t_0]\!]}$ affected by the exposure. Exposure and confounder profiles only affect the outcome through some summary variables $\mathscr{X}$ and $\mathscr{W}$. (b) Over-simplified causal model associated with the longitudinal model given in Figure 1.3 (a). (c) Other possibility of over-simplified causal model associated with the longitudinal model given in Figure 1.3 (a).

simplified causal models given in Figure 1.2 (b) or (c) would usually be considered to perform the analyses, depending on whether $(W_t)_t$ is mainly seen as a confounder or as a mediator.

The second situation is when the available data for the exposures corresponds to respective summary measures of their levels up to inclusion in the study. Consider for example the causal model given in Figure 1.3 (a), where the exposure of interest $(X_t)_t$ could again stand for the obesity status at different ages, $(W_t)_t$ could include alcohol intake, physical activity and diet at different ages, and $\mathscr{X}$ and $\mathscr{W}$ would be appropriate summary measures of $\bar{X}_{t_0}$ and $\bar{W}_{t_0}$, respectively; then the analyses would probably be performed under the over-simplified causal model given in Figure 1.3 (b) or (c).

In each of these two situations, the conditions guaranteeing a clear interpretation of the causal effects estimated under over-simplified models are very restrictive and, overall, our results emphasize the need for repeated measurements to perform valid causal analyses.

1.2.3 Probabilistic models for dimension-reduction

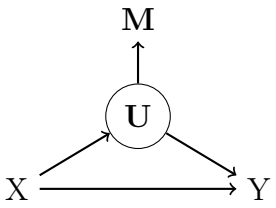
High dimensional mediation analysis and latent variable models

Omics data, and notably metabolomics data, provide important opportunities for the investigation of biological mechanisms possibly involved in a given exposure-cancer relationships, using an agnostic approach. However, their analysis is quite challenging because of their dimensionality and complex structure. Consider for instance the case-control studies on the endometrial cancer nested within the EPIC cohort, where targeted metabolomics data describing the levels of ~ 150 metabolites have been acquired through the BIOCRATES kit. Such data set is of relatively high dimension and possibly contains redundant information, as some metabolites are highly correlated with each other. Recently, methods have been introduced to allow the practicable application of high dimensional mediation analysis, including an extension of interventional direct and indirect

effects to a setting with a high-dimensional set of mediators (Loh et al., 2020), as well as a multiple testing procedure to perform selection on the set of potential mediators (Sampson et al., 2018), and also a regularized model of structural equations (Zhao and Luo, 2016). On the other hand, the high-dimensionality and large correlations among the metabolites could be "simultaneously" tackled by reduction-dimension techniques. In particular, they could lead to the identification of a small number of uncorrelated metabolic components, or signatures, defined as linear combinations of the original metabolites, and summarizing the information contained in the whole set of metabolites, that could mediate the effect of the exposure on the outcome.

Principal Component Analysis (PCA) (Hotelling, 1933, Jolliffe, 2002) is among the most popular multivariate methods for dimension-reduction. Applied to the set of 150 metabolites, it would identify mutually orthogonal components with maximal variances, defined as linear combinations of the metabolites. These principal components would then constitute the metabolic signatures. Partial Least Squares (PLS) Regression (Wold, 1985) is another popular multivariate dimension-reduction technique, which further allows to include additional information to define the metabolic signatures, for instance from a continuous exposure such as Body Mass Index (BMI). In that case, PLS Regression would also identify mutually orthogonal component defined as linear combinations of the metabolites, but the weights involved in the combinations would be chosen so that the components have maximal covariance with BMI. This approach has been applied in cancer epidemiology (Assi et al., 2015a). In a similar way, PLS Discriminant Analysis (Barker and Rayens, 2003), a variant of PLS Regression that allows the inclusion of information from a categorical variable, could be used to obtain metabolic signatures associated with cancer occurrence. However, these strategies are actually not perfectly suited to the mediation analysis setting. Indeed, PCA focuses on the set of metabolites only, so that the PCA signatures are not necessarily associated with the exposure, nor with the outcome. On the other hand with the PLS Regression (resp. PLS-DA), metabolic signatures of BMI (resp. cancer risk), can be obtained, but they are not necessarily associated with the outcome (resp. the exposure). In this regard, several methods have been introduced to account for the fact, in the mediation analysis setting, the signatures should be associated with both the exposure and the outcome. Geuter et al. (2020) proposed a method that look for signatures that maximize the indirect effect between the exposure and outcome. Alternatively, Chén et al. (2018) and Derkach et al. (2019) proposed to used latent variable models, where signatures can be estimated via (penalized) likelihood maximization.

The idea of using latent variable models to perform reduction-dimension is not new. For instance, Tipping and Bishop (1999) have used a Gaussian latent variable model to propose a probabilistic formulation of PCA (PPCA), and probabilistic formulations of PLS Regression have also been proposed by Li et al. (2015) and Zheng et al. (2016). The models proposed respectively by Chén et al. (2018) and Derkach et al. (2019) can be



Latent variable model for mediation analysis proposed by Derkach et al. (2019), in the case of a continuous outcome:

$$\mathbf{U} = X\delta^\top + \varepsilon_U,$$
$$\mathbf{M} = \mathbf{U}B^\top + \varepsilon_M\ , \qquad Y = Xb + \mathbf{U}\gamma + \varepsilon_Y.$$

Under the assumptions:

(1) $X \sim \mathcal{N}(0, \sigma_X^2)$.
(2) $\varepsilon_U \sim \mathcal{N}(\mathbf{0}_r; I_r)$. (3) $\varepsilon_M \sim \mathcal{N}(\mathbf{0}_{p_M}; \Psi_M)$.
(4) Ψ_M is a $p_M \times p_M$ diagonal matrix.
(5) $\varepsilon_Y \sim \mathcal{N}\big(0, \sigma_Y^2\big)$.
(6) $r < p_M$.

(a) (b)

Figure 1.4: (a) Graphical representation of the latent variable models for mediation analysis, proposed by Derkach et al. (2019) and Chén et al. (2018). Circled nodes represent sets of latent variables. (b) The latent variable model for mediation analysis proposed by Derkach et al. (2019), which can be depicted graphically as in Figure 1.4 (a).

depicted graphically as in Figure 1.4 (a), where three sets of observed variables X, $\mathbf{M}$ and Y are involved, as well as $\mathbf{U}$, a set of latent variables. They are more precisely defined by a system of structural equations involving the four sets of variables, together with a set of assumptions on the model parameters and on the distributions of some of the random variables involved in the model; see for instance the latent variable model for mediation analysis proposed by Derkach et al. (2019) in the case of a continuous outcome, in Figure 1.4 (b).

However, these models still need to be extended to study mechanisms underlying the obesity-cancer relationship, or more generally the lifestyle-cancer relationships. Indeed, several indicators can be used to define obesity, including BMI, but also the waist circumference, the waist-hip ratio, etc. Similarly, lifestyle encompasses a number of factors related to diet, physical activity, smoking status and intensity, alcohol intake, etc. In other words, epidemiologists now tend to consider multiple factors simultaneously and work with multivariate exposures. Then, dimension reduction needs to be performed for both the exposures and metabolites, while ensuring that the exposure signatures and metabolic signatures are associated with each other, and with cancer risk. In addition, even if the set of exposures is likely to be of low to moderate dimension, the set of mediators, e.g. metabolites, is usually of much larger dimension. Then, some level of regularization may be needed to encourage, e.g., sparsity in the weight vectors used to define the metabolic signatures. Under the model of Derkach et al. (2019) (in the case of a univariate exposure), the authors proposed to use an adaptive lasso penalty.

Contributions

We decided to extend the latent variable model proposed by Derkach et al. (2019) to a framework where the exposure is multivariate and where two sets of latent variables are present. More precisely, our estimation procedure used L_1-penalized versions of the likelihood to enforce sparsity in the weight vectors used for the construction of the metabolic signatures. As the model proposed by Derkach et al. (2019) is inspired by the Factor Analysis, the weight vectors are identifiable only up to an orthogonal transformation; for this reason, we decided to use slightly different constraints on the model parameters and distributions of the variables involved in the model compared to the ones used by Derkach et al. (2019). More precisely, our choice was inspired by the constraints used by el Bouhaddani et al. (2018) for their Probabilistic Partial Least Squares (PPLS) model, which guarantee the identifiability of the model parameters (up to sign for some of them).

However, when studying the identifiability of our latent variable model for high-dimensional mediation analysis, we noticed a severe defect: the constraints on the model parameters are too strong, and the parameters of interest then reduce to parameters that could be obtained under much simpler models. More precisely, our model defines a subset of very particular distributions for the three sets of observed variables (exposures, metabolites, and outcomes), where the weight vectors to be used for the construction of the exposure and metabolic signatures could be obtained from two PCAs or PPCAs, run separately on the exposure and metabolite sets. Of course, this greatly limits the applicability and interest of our model, although it looked specifically tailored for mediation analysis at first glance. As a matter of fact, we noticed that several other models proposed in the literature suffer from similar defects, including the model proposed by Derkach et al. (2019) for mediation analysis, and the PPLS model proposed by el Bouhaddani et al. (2018). In Chapter 4, and for simplicity, we focus on the later model under which two sets of observed variables only are considered (e.g., exposures and metabolites), to precisely describe this limitation. More precisely, we study the PPLS model proposed by el Bouhaddani et al. (2018), and show that this model defines a very particular subset of distributions too, under which the components (signatures) with maximal covariance are necessarily of maximal variances as well. We further illustrate this limitation through simulated examples, and propose a simple extension, which corrects the defect of the initial model.

1.2.4 The use of cross-validation for the calibration of the tuning parameter in the adaptive lasso

Finally, the study and the use of L_1-norm penalization for our initial model for high-dimensional mediation analysis has further led us to identify a defect of the K-fold cross-validation when applied for the calibration of the regularization parameter in the adaptive

lasso the adaptive lasso (Bühlmann and Meier, 2008, Zou, 2006). Again, this defect was not specific to our model, and is already present in simpler models, such as linear regression models. Consider a linear regression model of the form $Y = X\boldsymbol{\beta}^* + \xi$, where $Y = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ is the outcome vector, $X = (\mathbf{x}_1, ..., \mathbf{x}_n)^T \in \mathbb{R}^{n \times p}$ is the design matrix, $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*) \in \mathbb{R}^p$ is the vector of parameters to be estimated, and $\xi \in \mathbb{R}^n$ is some random noise. For any given vector of non-negative weights $\mathbf{w} = (w_j)_{1 \leq j \leq p}$ and any value of the regularization parameter $\lambda \geq 0$, the adaptive lasso estimator $\hat{\boldsymbol{\beta}}_{\text{ada}}(\lambda, \mathbf{w})$ is defined as any minimizer over $\boldsymbol{\beta} \in \mathbb{R}^p$ of $\dfrac{||Y - X\boldsymbol{\beta}||_2^2}{n} + \lambda \sum_{j=1}^p w_j |\beta_j|$ (Tibshirani, 1996).

The adaptive lasso is an extension of the standard lasso, which corresponds to the particular case where all the w_j's are set to 1. In the adaptive lasso, the weights w_j typically rely on some initial estimates of the parameters $(\beta_j^*)_{1 \leq j \leq p}$, and popular versions of the adaptive lasso include the original adaptive lasso introduced by Zou (2006), where weights are derived from Ordinary Least Squares (OLS) estimates, as well the version proposed by Bühlmann and Meier (2008), namely the one-step lasso, where weights are computed from initial lasso estimates. The adaptive lasso is very popular in practice as it can be solved very efficiently, and was shown to generally outperform the standard lasso; see Zou (2006), Bühlmann and van De Geer (2011).

Irrespective of the particular choice of the weights, the tuning parameter λ controls the amount of regularization through the weighted L_1-norm. In practice, an appropriate value for this parameter has to be selected to guarantee that $\hat{\boldsymbol{\beta}}_{\text{ada}}(\lambda, \mathbf{w})$ has good statistical performance. A popular strategy for the calibration of the tuning parameter relies on K-fold cross-validation (Hastie et al., 2009), whose principle can be summarized as follows. Denote by $\Lambda = (\lambda_1, \ldots, \lambda_R)$ a sequence of candidate values for the tuning parameter. First, the original sample $D = (y_i, \mathbf{x}_i)_{1 \leq i \leq n}$ is split into $K \geq 2$ balanced folds $D^{(1)}, \ldots, D^{(K)}$, with $D = \cup_{k=1}^K D^{(k)}$. It then consists of K steps: at each step k, (i) the fold $D^{(k)}$ is used as an "independent" test sample, while the remaining $K - 1$ folds $D \setminus D^{(k)}$ are combined and jointly used as the training sample, (ii), for every $r = 1, \ldots, R$, the adaptive lasso estimator $\hat{\boldsymbol{\beta}}_{\text{ada}}^{(k)}(\lambda_r, \mathbf{w})$ is estimated on the training sample $D \setminus D^{(k)}$, and (iii) its prediction error, $\text{PredErr}(\text{D}^{(\text{k})}, \hat{\boldsymbol{\beta}}_{\text{ada}}^{(\text{k})}(\lambda_{\text{r}}, \mathbf{w}))$ is evaluated on the test sample $D^{(k)}$. The cross-validated prediction error is finally defined as the average of these K prediction errors $\frac{1}{K} \sum_{k=1}^K \text{PredErr}(\text{D}^{(\text{k})}, \hat{\boldsymbol{\beta}}_{\text{ada}}^{(\text{k})}(\lambda_{\text{r}}, \mathbf{w}))$, and the optimal tuning parameter is selected as the one minimizing this cross-validated error.

The overall principle of cross-validation is to mimic "independent" test samples. In particular, for the K-fold cross-validation to perform well, at each step k the whole estimation procedure should be performed on the training sample $D \setminus D^{(k)}$, and should not use any information from the test sample $D^{(k)}$. Yet, the weights used in the adaptive lasso are derived from initial estimates computed on the entire original sample D. Therefore, considering the estimation of the adaptive lasso estimator as a whole, it does use information from the test samples. Under the simple setting of linear regression models, we

show empirically in Chapter 5, for several standard choices of the weights, that the K-fold cross-validation is not suitable for the calibration of the tuning parameter in the adaptive lasso. We propose a simple alternative cross-validation scheme to rectify the defect, which is further shown to outperform the standard K-fold cross-validation on simulated examples.

Chapter 2

Which practical interventions does the *do*-operator refer to in causal inference? Illustration on the example of obesity and cancer.

This Chapter corresponds to the preprint available at https://arxiv.org/abs/1901.00772, and written with Vivian Viallon.

Abstract

For exposures X like obesity, no precise and unambiguous definition exists for the hypothetical intervention $do(X = x_0)$. This has raised concerns about the relevance of causal effects estimated from observational studies for such exposures. Under the framework of structural causal models, we study how the effect of $do(X = x_0)$ relates to the effect of interventions on causes of X. We show that for interventions focusing on causes of X that affect the outcome through X only, the effect of $do(X = x_0)$ equals the effect of the considered intervention. On the other hand, for interventions on causes W of X that affect the outcome not only through X, we show that the effect of $do(X = x_0)$ only partly captures the effect of the intervention. In particular, under simple causal models (e.g., linear models with no interaction), the effect of $do(X = x_0)$ can be seen as an indirect effect of the intervention on W.

2.1 Introduction

Because most epidemiological results are derived from observational data, their causal interpretation has always been at the center of concern (Rothman et al., 2008). Causal inference theory, which has attracted a lot of interest in the last few decades, has proved

useful to formally describe conditions ensuring the causal validity of results derived from observational data (Glymour and Greenland, 2008, Hernán and Robins, 2020, Pearl, 2000, Rothman and Greenland, 2005, Rubin, 1974). For example, a number of sets of sufficient conditions has been established for the identifiability of causal effects in the presence of confounding or non-random selection. Under the so-called Structural Causal Models (SCMs) (Pearl, 1995, 2000), and further assuming that the structure of the underlying Directed Acyclic Graph (DAG) is known, a key condition for the identifiability of the causal effect is exchangeability, or ignorability (Hernán and Robins, 2020, Pearl, 2000, Rosenbaum and Rubin, 1983). In particular, ignorability has been shown to hold conditionally on any set of variables satisfying the back-door criterion (Pearl, 1993, 2000). Then, a variety of statistical approaches have been proposed for the estimation of causal effects under increasingly complex settings including time-varying confounding, failure time data, etc. Among other approaches, we shall mention the parametric g-formula, inverse probability weighting approaches, g-estimation and doubly robust procedures (Hernán and Robins, 2020, Lunceford and Davidian, 2004, Pearl, 2000).

Even if their use has been controversial (Dawid, 2000), counterfactual variables, or potential outcomes, are key to most causal inference theories commonly considered nowadays, in epidemiology, social science, statistics and computer science. The *do*-calculus that accompanies SCMs allows precise definitions of these variables and their joint distribution (Pearl, 2000). Here, we will use the notation $Y^{(X=x_0)}$ to denote the counterfactual variable representing the outcome that would have been observed in the counterfactual world $\Omega^{(X=x_0)}$ that would have followed the hypothetical intervention $do(X = x_0)$, where X is the exposure of interest and x_0 is any potential value for this exposure (Pearl, 1995, 2000, Rubin, 1974). For simplicity, we will focus on binary outcomes, and we let $\mathbb{P}(Y = 1|do(X = x_0)) = \mathbb{P}(Y^{(X=x_0)} = 1)$ denote the probability of observing the outcome in this counterfactual world.

For some exposures, the lack of a precise and unambiguous definition for the intervention $do(X = x_0)$ has raised some concerns in the literature (Cole and Frangakis, 2009, Hernán, 2016, Hernán and Taubman, 2008, Hernán and VanderWeele, 2011, Pearl, 2010, Petersen, 2011, Petersen and van der Laan, 2014, van der Laan et al., 2005, Vandenbroucke et al., 2016, VanderWeele and Hernán, 2013). For example, consider the case where X stands for a binary variable indicating obesity status at 20 years of age. In a population of lean teenagers, or even newborns, the hypothetical intervention $do(X = x_0)$, for $x_0 = 0$ (or $x_0 = 1$), could then correspond to a typically adaptive and dynamic intervention that would ensure that individuals stay lean (or get obese) by the age of 20. However, these interventions are not well-defined, in the sense that different "versions" may lead to the same obesity value x_0 at 20 years-old. For instance, in the "stay lean" arm ($do(X = 0)$), individuals may be asked to do 45 minutes of physical exercise a day, or 72 minutes of physical exercise a day. They could also be asked to adhere to a healthy diet, etc. In

addition, some of the versions ensuring that $X = 0$ at 20 years old may be impossible to apply in practice, such as those involving genetic factors.

More generally, this situation of a treatment with different versions, or compound treatment, violates the "no-multiple-versions-of-treatment assumption", which is part of the "Stable Unit Treatment Value Assumption" (SUTVA) (Rubin, 1980, VanderWeele and Hernán, 2013). This has led to some debate around the relevance, for public health matters, of the causal effects estimated from observational studies in such cases. Interestingly, most arguments have been based by considering the situation where "treatment precedes versions of that treatment", while situations where "versions precede treatment" were only quickly mentioned, if at all (Hernán and VanderWeele, 2011, Petersen, 2011, VanderWeele and Hernán, 2013). Here, we consider the situations where versions precede treatment, in which case these versions can be seen as particular levels for the causes of X. Then, focusing on situations where direct interventions on X are impractical, we inspect how the effect of the hypothetical intervention $do(X = x_0)$ relates to the effects of interventions on causes of X. We show that the effect of the hypothetical intervention $do(X = x_0)$ equals the effect of particular interventions on causes of X that are causes of Y through X only, as expected. However, for causes W that influence Y not only through X, the causal effect of X differs from the causal effect of interventions on W. For example, in the particular case of obesity and cancer occurrence, the effect of $do(X = x_0)$ is different from the effects of interventions on diet or physical activity, except for cancers whose risk is not directly associated with diet and/or physical activity.

To make our illustrative example even more concrete, we assume throughout that we intend to estimate the causal effect of obesity at 20 years of age on the occurrence of cancer by the age of 50. A typical prospective cohort study would sample individuals who are cancer-free at the age of 20, record information regarding their obesity status and other variables (potential confounders, etc.) at inclusion, follow these individuals over the age interval 20-50 and finally record cancer occurrence by the age of 50. Denote by $X \in \{0, 1\}$ and $Y \in \{0, 1\}$ the binary variables representing obesity at 20 and cancer occurrence between 20 and 50. For simplicity, we further assume the absence of competing events and censoring.

The rest of the article is organized as follows. Even if this is highly unlikely in our illustrative example, we start by considering the unconfounded setting where all causes of X are causes of Y through X only. Then, in Section 2.3, we consider a more realistic setting where confounders are present. We shall stress that this second setting is still an over-simplified version of the causal model in our illustrative example (see the Discussion 2.4). Yet, we believe it is instructive to describe the relationship between the intervention $do(X = x_0)$ and its multiple versions. Under both settings, we consider the situation where some causes are modifiable, while others are not. Section 2.4 presents some concluding remarks and discussion. Proofs of our main results are presented in Appendix 2.A and

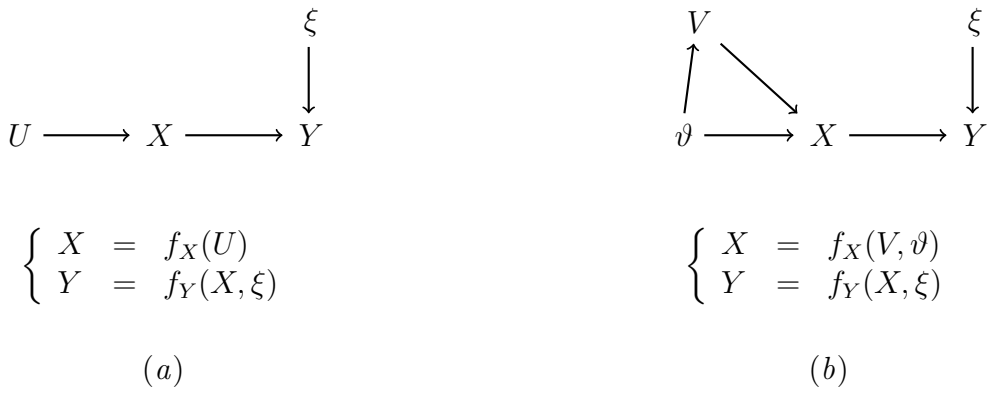


Figure 2.1: DAGs and associated structural equations in the unconfounded case: (a) Standard causal model, without confounding. (b) Decomposing U as $U = (V, \vartheta)$, where V and ϑ correspond to modifiable and non-modifiable causes of X, respectively.

Appendix 2.B.

2.2 The unconfounded case

Because exposure X is not randomized in our prospective cohort study, identifiability of the causal effect of X on Y is generally not guaranteed. A particular situation when this causal effect is identifiable is when all causes of X, denoted by U in this simple case, are causes of Y through X only. Even if this absence of confounders is highly unlikely in our illustrative example, it is instructive to consider this simple situation as a starting point. The more general situation where confounding is present is deferred to Section 2.3.

2.2.1 Preliminary derivations

Consider that the data available in our cohort study are generated by a causal model with associated DAG and structural equations as presented in Figure 2.1 (a). Variables ξ and U represent all causes of Y and X, respectively, and are assumed to be independent to each other. Both ξ and U may include purely random components. Given the structural equations attached to this simple causal model, we have $\{X = x\} \Rightarrow \{Y = Y^{(x)}\}$, so that consistency holds. Moreover, under this causal model, the ignorability condition $(Y^{(x)} \perp\!\!\!\perp X)$ holds. Then

$$\begin{aligned} ATE &= \mathbb{P}(Y = 1|do(X = 1)) - \mathbb{P}(Y = 1|do(X = 0)) \\ &= \mathbb{P}(Y = 1|X = 1) - \mathbb{P}(Y = 1|X = 0), \end{aligned}$$

and the causal effect of X on Y is identifiable. But, when direct interventions on X are impractical, and only interventions on the causes of X are practical, a natural question is the meaning of the hypothetical intervention $do(X = x)$. Consider the structural equation pertaining to exposure, $X = f_X(U)$, and set $f_X^{-1}(x_0) = \{u : f_X(u) = x_0\}$.

Of course, we have $X = x_0 \Leftrightarrow U \in f_X^{-1}(x_0)$. As a result, for any $u_{x_0} \in f_X^{-1}(x_0)$, $\mathbb{P}(Y = 1|do(U = u_{x_0})) = \mathbb{P}(Y = 1|do(X = x_0))$; see Appendix 2.A. In this simple case, all interventions $do(U = u_{x_0})$ on the causes of X which would yield $X = x_0$ share the same effect on Y: versions are irrelevant (Hernán and VanderWeele, 2011, VanderWeele and Hernán, 2013), and the causal effect $\mathbb{P}(Y = 1|do(X = x_0))$ estimated on the cohort is an estimate of this shared effect.

2.2.2 Distinguishing modifiable and non-modifiable causes

To gain insight from a practical standpoint, the previous analysis can be slightly refined by decomposing causes of X as $U = (V, \vartheta)$ where V and ϑ correspond to sets of modifiable and non-modifiable causes of X, respectively. See Figure 2.1 (b). Because non-modifiable causes may affect modifiable ones, while the former are unlikely to be affected by the latter, we do not consider the possibility of an arrow pointing from V to ϑ in Figure 2.1 (b). Causes ϑ are non-modifiable and the only interventions that could be practically set up are those on V. Denote the set of possible values of ϑ by $\mathscr{V}$. Then, for any $x \in \{0, 1\}$ and $\nu \in \mathscr{V}$, set $f_{X|\vartheta}^{-1}(x; \nu) = \{v : f_X(v, \nu) = x\}$. First assume that this set is non-empty for any $x \in \{0, 1\}$ and $\nu \in \mathscr{V}$: in other words, first assume that, for any $x \in \{0, 1\}$, and for any value ν for the non-modifiable factors ϑ, there exists some value v of the modifiable factors V such that $f_X(\nu, v) = x$. Now, for individuals such that $\vartheta = \nu_0$, for any $\nu_0 \in \mathscr{V}$, we have $X = x_0 \Leftrightarrow V \in f_{X|\vartheta}^{-1}(x_0; \nu_0)$. Therefore $\mathbb{P}(Y^{(V=v_{x_0}(\nu_0))} = 1|\vartheta = \nu_0) = \mathbb{P}(Y = 1|do(V = v_{x_0}(\nu_0)), \vartheta = \nu_0) = \mathbb{P}(Y = 1|do(X = x_0))$ for any $v_{x_0}(\nu_0) \in f_{X|\vartheta}^{-1}(x_0; \nu_0)$. Denote by $do(V = v_{x_0}(\vartheta))$ any intervention which sets, for all individuals in the population, the value of V according to the value ν_0 of ϑ, in such a way that for any individual with $\vartheta = \nu_0$, the intervention $do(V = v_{x_0}(\vartheta))$ sets V to $v_{x_0}(\nu_0) \in f_{X|\vartheta}^{-1}(x_0; \nu_0)$. Then, we have $\mathbb{P}(Y = 1|do(V = v_{x_0}(\vartheta))) = \mathbb{P}(Y = 1|do(X = x_0))$. In other words, versions are again irrelevant and any such intervention has the same effect on Y, which is $\mathbb{P}(Y = 1|do(V = v_{x_0}(\vartheta)) = \mathbb{P}(Y = 1|do(X = x_0))$.

Of course, unless there exists at least one value $v_1 \in \cap_{\nu \in \mathscr{V}}\{f_{X|\vartheta}^{-1}(x_0; \nu)\}$, only a dynamic, i.e. individual-specific, treatment can be adopted to attain this effect. For instance, consider the "stay lean" arm of the clinical trial mentioned in the Introduction 2.1. Because individuals may be more or less genetically predisposed to obesity, some individuals will have to make little effort to stay lean by the age of 20, while others will have to adopt a drastic diet and/or have intense physical activity, etc. We may stress that this heterogeneity among individuals is at the core of personalized (preventive) medicine and need to be acknowledged, rather than discarded, in causal inference. Similarly, our cohort reflects this heterogeneity: individuals sharing the same obesity status $\{X = x_0\}$, for $x_0 \in \{0, 1\}$, can differ regarding V and ϑ. More precisely, for $x_0 \in \{0, 1\}$, set $\mathscr{V}(x_0) = \{\nu \in \mathscr{V} : f_{X|\vartheta}^{-1}(x_0; \nu) \neq \emptyset\}$. The lean and obese groups in our

$$\begin{cases} X &= f_X(U, W) \\ Y &= f_Y(X, W, \xi) \end{cases} \qquad \begin{cases} X &= f_X(V, \vartheta, W, Z) \\ Y &= f_Y(X, W, Z, \xi) \end{cases}$$

(a) (b)

Figure 2.2: DAGs and associated structural equations in the presence of confounders (a) Standard causal model with confounding. (b) Distinguishing modifiable and non-modifiable causes of X in the presence of confounding.

cohort are sampled from

$$\{X = x_0\} = \bigcup_{\nu \in \mathscr{V}(x_0)} \left\{ \{\vartheta = \nu\} \bigcap \{V \in f^{-1}_{X|\vartheta}(x_0; \nu)\} \right\}$$

for $x_0 = 0$ and $x_0 = 1$, respectively. Again, if the model of Figure 2.1 (b) is correct, versions of the compound treatment obesity are not relevant (Hernán and VanderWeele, 2011, VanderWeele and Hernán, 2013). Therefore, how the levels of the causes of "obesity at 20 years of age" are mixed up in the group of obese, or lean, individuals in our cohort is not relevant either: our cohort would return unbiased estimates for the quantity $\mathbb{P}(Y = 1|do(X = x_0)) = \mathbb{P}(Y = 1|X = x_0)$, just as the clinical trial would. Then, the effect of the intervention $do(X = x_0)$ can again be interpreted as the effect of any intervention on the causes of X ensuring $X = x_0$.

If, for some x, there exist some values $\nu_1 \in \mathscr{V}$ of the non-modifiable variables ϑ such that the set $f^{-1}_{X|\vartheta}(x; \nu_1)$ is empty, the intervention $do(X = x)$ is purely theoretical for individuals such that $\vartheta = \nu_1$ since no practical intervention could yield $X = x$ for them. However, under the assumptions of SCMs, and if the DAG of Figure 2.1 (b) is correct, the effect of the hypothetical intervention $do(X = x_0)$ can still be estimated from our cohort study even if no practical intervention ensuring $X = x_0$ exists for individuals with $\vartheta = \nu_1$ (whenever the positivity condition further holds $(0 < \mathbb{P}(X = x_0) < 1)$). Indeed, we have $\mathbb{P}(Y = 1|do(X = x_0), \vartheta = \nu_1) = \mathbb{P}(Y = 1|do(X = x_0) = \mathbb{P}(Y = 1|X = x_0)$.

2.3 The more standard case with confounders

2.3.1 Preliminary analyses

We now turn our attention to the more common situation where confounding is present. Without loss of generality, assume that causes of X are grouped in two sets, W and U.

Here, and as above, causes in U are assumed to have an effect on Y through X only, while W is the set of common causes of X and Y, that is the set of confounders in the X-Y relationship. In our illustrative example, W could include gender, physical activity and dietary habit, while U might include genetic predisposition to obesity. Figure 2.2 (a) depicts the corresponding causal model. Assume for ease of notation that the set $\mathscr{W}$ of possible values for W is discrete. Further recall that consistency still holds, and assume that $0 < \mathbb{P}(X = 1|W = w) < 1$ for all w such that $\mathbb{P}(W = w) > 0$. Then, because $Y^{(x)} \perp\!\!\!\perp X|W$ under the model depicted in Figure 2.2 (a) , the causal effect of X on Y is identifiable. More precisely, we have

$$ATE = \sum_{w}[\mathbb{P}(Y = 1|X = 1, W = w) - \mathbb{P}(Y = 1|X = 0, W = w)]\mathbb{P}(W = w).$$

But, again, a natural question is how the hypothetical intervention $do(X = x)$ does relate to interventions on causes of X. Neglecting for now issues related to the possibility to apply these interventions in practice, these interventions can concern either (i) U only, (ii) W only, or (iii) both U and W.

First consider interventions on U only and set, for any $x \in \{0, 1\}$ and $w \in \mathscr{W}$, $f^{-1}_{X|W}(x; w) = \{u : f_X(u, w) = x\}$. For any $w_0 \in \mathscr{W}$, we have $X = x_0 \Leftrightarrow U \in f^{-1}_{X|W}(x_0; w_0)$ for individuals belonging to stratum $W = w_0$. Then, assume that $f^{-1}_{X|W}(x_0; w_0)$ is non-empty for all $(x_0, w_0) \in \{0, 1\} \times \mathscr{W}$ and denote by $do(U = u_{x_0}(W))$ any intervention setting U to any value $u_{x_0}(w_0) \in f^{-1}_{X|W}(x_0; w_0)$ for individuals in stratum $W = w_0$, for all $w_0 \in \mathscr{W}$. Arguing as in Section 2.2.2, we get $\mathbb{P}(Y = 1|do(U = u_{x_0}(W))) = \mathbb{P}(Y = 1|do(X = x_0))$; see Section 2.B.1 in the Appendix. Again, versions are irrelevant, and any such intervention has the same effect on Y, which is $\mathbb{P}(Y = 1|do(X = x_0))$.

Now consider interventions on W only and set, for any $x \in \{0, 1\}$ and $u \in \mathscr{U}$, $f^{-1}_{X|U}(x; u) = \{w : f_X(u, w) = x\}$. Then, assume that $f^{-1}_{X|U}(x; u)$ is non-empty for every $(x, u) \in \{0, 1\} \times \mathscr{U}$, and for any $u_0 \in \mathscr{U}$, denote by $w_{x_0}(u_0)$ one given element of $f^{-1}_{X|U}(x_0; u_0)$. Given this particular collection of values $(w_{x_0}(u))_{u\in\mathscr{U}}$, denote by $do(W = w_{x_0}(U))$ the intervention which sets W to $w_{x_0}(u_0)$ for individuals in stratum $U = u_0$, for all $u_0 \in \mathscr{U}$. Arguing as before, it comes that $\mathbb{P}(Y = 1|do(W = w_{x_0}(U))) = \mathbb{P}(Y = 1|do(X = x_0, W = w_{x_0}(U)))$, which generally differs from $\mathbb{P}(Y = 1|do(X = x_0))$. The intervention $do(W = w_{x_0}(U))$ does entail $X = x_0$ for all individuals, but because W has an effect on Y not only through X, the effect of $do(W = w_{x_0}(U))$ is not entirely captured by that of $do(X = x_0)$. Actually, X can be seen as a mediator in the $W - Y$ relationship, and, under simple models, in particular in the absence of interaction between X and W, the effect of $do(X = x_0)$ is actually related to the indirect effect of the intervention $do(W = w_{x_0}(U))$, through X; see Section 2.B.3 in the Appendix. It is also important to note that $\mathbb{P}(Y = 1|do(W = w_{x_0}(U))$ depends on the

collection of values $(w_{x_0}(u))_{u\in\mathscr{U}}$. If w_0 and $\tilde{w}_0$ are two distinct elements of $f^{-1}_{X|U}(x_0;u_0)$ for some $u_0 \in \mathscr{U}$, then $\mathbb{P}(Y=1|do(W=w_0),U=u_0) = \mathbb{P}(Y^{(W=w_0,X=x_0)}=1)$, while $\mathbb{P}(Y=1|do(W=\tilde{w}_0),U=u_0) = \mathbb{P}(Y^{(W=\tilde{w}_0,X=x_0)}=1)$. The difference between these two quantities is related to the direct effect of W, and reflects the fact that two interventions on W sharing the same effect on X do not necessarily have the same effects on Y when W has a direct effect on Y: in this case, versions of the compound treatment are relevant.

Now, if $f^{-1}_{X|U}(x;u)$ is empty for some $(x,u) \in \{0,1\}\times\mathscr{U}$, then no intervention on W only can ensure $X=x$ for individuals in stratum $U=u$. Similarly, if $f^{-1}_{X|W}(x;w)$ is empty for some pair (x,w), then no intervention on U only can ensure $X=x$ for individuals in stratum $W=w$. Then, consider interventions on both W and U, and set $f^{-1}_X(x) = \{(w,u) : f_X(u,w)=x\}$. For any $(w_0,u_0)\in f^{-1}_X(x_0)$, it is easy to show that $\mathbb{P}(Y=1|do(W=w_0,U=u_0)) = \mathbb{P}(Y^{(W=w_0,X=x_0)}=1)$. Therefore, interventions on both W and U that ensure $X=x_0$ are similar to interventions on W only: their effects are generally not uniquely defined (they depend on the particular pair of values $(w_0,u_0)\in f^{-1}_X(x_0)$) and only partly capture the effect of interventions on X.

2.3.2 Distinguishing modifiable and non-modifiables causes

All the analyses above can be refined by acknowledging that some causes in U and W are modifiable, while others are not, and by considering interventions on modifiable causes only. See Figure 2.2 (*b*). Compared to Section 2.3.1, notations become a little more complex, but conclusions remain mostly similar. For instance, consider interventions on both V and W, where V is a modifiable cause of X with no direct effect on Y, while W is a modifiable confounder in the $X-Y$ relationship. For any $x_0\in\{0,1\}$ and any potential values ν and z for non-modifiable causes ϑ and Z, assume that the set $f^{-1}_{X|\vartheta,Z}(x_0;\nu,z) = \{(v,w) : f_X(v,\nu,w,z)=x_0\}$ is non-empty, and denote by $(v_{x_0}(\nu,z),w_{x_0}(\nu,z))$ one given element in this set. Then denote by $do(V=v_{x_0}(\vartheta,Z),W=w_{x_0}(\vartheta,Z))$ the intervention setting V to $v_{x_0}(\nu_0,z_0)$ and W to $w_{x_0}(\nu_0,z_0)$ for any individuals in stratum $\{\vartheta=\nu_0\}\cap\{Z=z_0\}$, for all ν_0,z_0. Arguing as before, it can be shown that $\mathbb{P}(Y=1|do(V=v_{x_0}(\vartheta,Z),W=w_{x_0}(\vartheta,Z))) = \mathbb{P}(Y=1|do(X=x_0,W=w_{x_0}(\vartheta,Z)))$. This quantity generally differs from $\mathbb{P}(Y=1|do(X=x_0)$ and the reason again is that the intervention $do(V=v_{x_0}(\vartheta,Z),W=w_{x_0}(\vartheta,Z))$ not only ensures that $X=x_0$, but it also has a direct effect on Y through the intervention on W.

2.4 Conclusion-Discussion

In this article, we showed how the hypothetical intervention $do(X=x_0)$, when impossible to apply in practice, relates to interventions on causes of X. Basing our arguments on structural causal models, our conclusions are in line with those of Petersen (2011): the DAG which represents our assumptions on the causal model under study is basically

sufficient to precisely understand how $do(X = x_0)$ can be interpreted. When interventions on causes of X that are causes of Y through X only exist, the effect of $do(X = x_0)$ captures the effect of such interventions. However, for causes of X, say W, that cause Y not only through X, the effect of $do(X = x_0)$ only partly captures the effect of interventions on W. Under simple causal models, the effect of $do(X = x_0)$ is related to the indirect effect of interventions on W.

Taking the example of obesity (at 20 years old) and the risk of cancer (by the age of 50), our results confirm concerns raised by several authors (Hernán and Taubman, 2008, Hernán and VanderWeele, 2011, VanderWeele and Hernán, 2013): because most modifiable causes of obesity can be regarded as confounders in the obesity-cancer relationship, the effect of obesity estimated from observational data likely differs from the effect of interventions on these causes, which could be estimated through clinical trials. At this point, however, we may insist on the fact that, if all modifiable causes of obesity are confounders in the obesity-cancer relationship, then clinical trials would not yield an estimate of the effect of obesity on cancer. Instead, a clinical trial would return an estimate of the causal effect of the considered intervention on cancer, and this effect would only partly capture the effect of obesity. Consider again the clinical trial sketched in the Introduction 2.1. More precisely, consider a randomized clinical trial where the study population, corresponding, e.g. to lean teenagers, is randomly assigned to two arms. Denote by U and Z the other, possibly non-modifiable, causes of X, with Z corresponding to common causes of Y and X, and U corresponding to causes of Y through X only. In this setting, observe that $Y^{X=x} \not\perp\!\!\!\perp W$ while $Y^{X=x} \perp\!\!\!\perp \{W, Z\}$ in general. Denote by $\mathscr{U}$ and $\mathscr{Z}$ the sets of possible values for U and Z, respectively. Then, an "ideal" clinical trial would consist in randomly assigning individuals to one of the following two groups: those for whom W would be set to $w_1(U, Z)$ and those for whom W would be set to $w_0(U, Z)$, for two given collections of values $(w_0(u, z))_{u\in\mathscr{U},z\in\mathscr{Z}}$ and $(w_1(u, z))_{u\in\mathscr{U},z\in\mathscr{Z}}$, where $w_0(u, z)$ and $w_1(u, z)$ ensure that $X = 0$ and $X = 1$, respectively, for individuals with $U = u$ and $Z = z$. Assuming complete compliance, and arguing as in Section 2.3, it is easy to show that the comparison of these two groups would return an estimate of the effect of this particular intervention on W, not that of X. Comparisons should be made between groups of individuals sharing the same value for W and Z to obtain a valid estimate of the effect of obesity, within strata defined by W and Z. In other words, under this ideal clinical trial setting, non-modifiable confounders in the $X - Y$ relationship would still have to be measured and controlled for to unbiasedly estimate the causal effect of obesity, within strata defined by W and Z. When controlled for a sufficient set of confounders, analyses based on observational studies can be used to derive unbiased estimates of these same effects.

There are a number of subtleties that we neglected for the sake of simplicity. First, a clinical trial whose objective is to prevent obesity by the age of 20 would typically

not only be dynamic, but also adaptive, i.e. the intervention is not only subject-specific, but it is also time-dependent. A good example is the Feeding Dynamic Intervention, to prevent childhood obesity[1]. Similarly, although we focused on time-fixed exposure and confounders, they are all time-varying in the population. For instance, physical activity and food intakes vary over the age interval $[0, 20)$, and the corresponding variables are all potential confounders in the relationship between obesity at 20 years-old and cancer occurrence before 50 years-old. Another important time-varying cause of obesity at 20 years-old is obesity over the age interval $[0, 19)$. Consequently, individuals in the two groups of our cohort, obese and lean at 20 years-old, do not only differ because of their status regarding obesity at 20 years of age, they also typically differ with respect to their histories regarding obesity, physical activity and dietary habits. This can lead to biases if these histories are not appropriately accounted for in the analysis (Etievant and Viallon, 2020). Second, selection bias may also be at play in our cohort study since only individuals who are cancer-free at 20 can be included. This selection bias will be more severe if cancer risk before 20 years old is associated to levels of obesity, physical activity and dietary habits over the age interval [0, 19]. This selection bias due to prevalent exposure and depletion of susceptibles has been put forward as one of the reasons explaining the discrepancies between results obtained through observational and interventional data when studying the association between hormone replacement therapy and coronary heart disease for instance (Hernán et al., 2008).

Disclaimers

Where authors are identified as personnel of the International Agency for Research on Cancer / World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer / World Health Organization.

2.A Appendix: Proof in the unconfounded case

Under the model depicted in Figure 2.1 (a), we have

$$\begin{aligned}\mathbb{P}(Y = 1|do(U = u_{x_0})) &= \mathbb{P}(Y^{(U=u_{x_0})} = 1) \\ &= \mathbb{P}(f_Y(X^{(U=u_{x_0})}, \xi) = 1) \\ &= \mathbb{P}(f_Y(x_0, \xi) = 1)\end{aligned}$$

[1] https://clinicaltrials.gov/ct2/show/NCT01515254

$$= \mathbb{P}(Y^{(x_0)} = 1)$$
$$= \mathbb{P}(Y = 1|do(X = x_0)).$$

2.B Appendix: Proof in the confounded case

2.B.1 Interventions of type (i)

Assume that $f^{-1}_{X|W}(x_0; w_0)$ is non-empty for any x_0, w_0. Then, under the model depicted in Figure 2.2 (a), we have, for any $u_{x_0}(w_0) \in f^{-1}_{X|W}(x_0; w_0)$

$$\begin{aligned}\mathbb{P}(Y = 1|do(U = u_{x_0}(w_0)), W = w_0) &= \mathbb{P}(Y^{(U=u_{x_0}(w_0))} = 1|W = w_0)\\ &= \mathbb{P}(f_Y(X^{(U=u_{x_0}(w_0))}, W, \xi) = 1|W = w_0)\\ &= \mathbb{P}(f_Y(x_0, w_0, \xi) = 1)\\ &= \mathbb{P}(Y^{(X=x_0,W=w_0)} = 1)\\ &= \mathbb{P}(Y = 1|do(X = x_0, W = w_0))\\ &= \mathbb{P}(Y = 1|do(X = x_0), W = w_0),\end{aligned}$$

where the last equality follows from rule 2 of the do-calculus(Pearl, 2000).

Moreover,

$$\begin{aligned}\mathbb{P}(Y = 1|do(U = u_{x_0}(W))) &= \sum_{w_0} \mathbb{P}(Y = 1|do(U = u_{x_0}(w_0)), W = w_0)\mathbb{P}(W = w_0)\\ &= \sum_{w_0} \mathbb{P}(Y = 1|do(X = x_0), W = w_0)\mathbb{P}(W = w_0)\\ &= \mathbb{P}(Y = 1|do(X = x_0)).\end{aligned}$$

2.B.2 Interventions of type (ii)

Assume that $f^{-1}_{X|U}(x_0; u_0)$ is non-empty for any x_0, u_0. Then, under the model depicted in Figure 2.2 (a), we have, for any $w_{x_0}(u_0) \in f^{-1}_{X|U}(x_0; u_0)$

$$\begin{aligned}\mathbb{P}(Y = 1|do(W = w_{x_0}(u_0)), U = u_0) &= \mathbb{P}(Y^{(W=w_{x_0}(u_0))} = 1|U = u_0)\\ &= \mathbb{P}(f_Y(X^{(W=w_{x_0}(u_0))}, w_{x_0}(u_0), \xi) = 1|U = u_0)\\ &= \mathbb{P}(f_Y(x_0, w_{x_0}(u_0), \xi) = 1|U = u_0)\\ &= \mathbb{P}(f_Y(x_0, w_{x_0}(u_0), \xi) = 1)\\ &= \mathbb{P}(Y^{(X=x_0,W=w_{x_0}(u_0))} = 1).\end{aligned}$$

2.B.3 Relationship with indirect effects

Denote by $(w_1(u_0), w_0(u_0))_{u_0 \in \mathscr{U}}$ two given collections of values such that $w_1(u_0) \in f^{-1}_{X|U}(1; u_0)$ and $w_0(u_0) \in f^{-1}_{X|U}(0; u_0)$. Further let $do(W = w_1(U))$ and $do(W = w_0(U))$ denote two given interventions setting W to $w_1(u_0) \in f^{-1}_{X|U}(1; u_0)$ and $w_0(u_0) \in f^{-1}_{X|U}(0; u_0)$, respectively, for individuals in stratum $U = u_0$, for all $u_0 \in \mathscr{U}$. We have

$$\begin{aligned}
\mathbb{E}(Y^{(w_1(U))} - Y^{(w_0(U))}) &= \sum_u \mathbb{E}(Y^{(w_1(u))} - Y^{(w_0(u))}|U = u)\mathbb{P}(U = u) \\
&= \sum_u \mathbb{E}(Y^{(w_1(u), X^{(w_1(u))})} - Y^{(w_0(u), X^{(w_0(u))})}|U = u)\mathbb{P}(U = u) \\
&= \sum_u \{\mathbb{E}(Y^{(w_1(u), X^{(w_1(u))})} - Y^{(w_1(u), X^{(w_0(u))})}|U = u) \\
&\quad + \mathbb{E}(Y^{(w_1(u), X^{(w_0(u))})} - Y^{(w_0(u), X^{(w_0(u))})}|U = u)\}\mathbb{P}(U = u) \\
&= \sum_u \mathbb{E}(Y^{(w_1(u), x_1)} - Y^{(w_1(u), x_0)} + Y^{(w_1(u), x_0)} - Y^{(w_0(u), x_0)})\mathbb{P}(U = u).
\end{aligned}$$

The term $\sum_u \mathbb{E}(Y^{(w_1(u), x_1)} - Y^{(w_1(u), x_0)})\mathbb{P}(U = u)$ can be regarded as an indirect effect since the level of W is held fixed and only the value of X changes from x_0 to x_1 which, for individuals in stratum $U = u$, equal $X^{(W = w_0(u))}$ and $X^{(W = w_1(u))}$ respectively. More precisely, we have

$$\begin{aligned}
&\sum_u \mathbb{E}(Y^{(w_1(u), x_1)} - Y^{(w_1(u), x_0)})\mathbb{P}(U = u) \\
&\quad = \sum_u \{\mathbb{E}(Y|W = w_1(u), X = x_1) - \mathbb{E}(Y|W = w_1(u), X = x_0)\}\mathbb{P}(U = u).
\end{aligned}$$

Under the model depicted in Figure 2.2 (a), recall we have

$$\begin{aligned}
&\mathbb{E}(Y|do(X = x_1)) - \mathbb{E}(Y|do(X = x_0)) \\
&\quad = \sum_w \{\mathbb{E}(Y|W = w, X = x_1) - \mathbb{E}(Y|W = w, X = x_0)\}\mathbb{P}(W = w).
\end{aligned}$$

Under simple causal models, for instance when $f_Y(W, X, \xi) = \alpha^T W + \beta X + \xi$, the two quantities, $\sum_u \mathbb{E}(Y^{(w_1(u), x_1)} - Y^{(w_1(u), x_0)})\mathbb{P}(U = u)$ and $\mathbb{E}(Y|do(X = x_1)) - \mathbb{E}(Y|do(X = x_0))$, coincide and equal β. However, under more complex models, these two quantities are typically different. Even under linear models, if interaction terms of the form $\gamma^T WX$ are present in function f_Y, these two terms are typically different and $\sum_u \mathbb{E}(Y^{(w_1(u), x_1)} - Y^{(w_1(u), x_0)})\mathbb{P}(U = u)$ would actually depend on the collection of values $\{w_1(u), u \in \mathscr{U}\}$.

Chapter 3

Causal inference under over-simplified longitudinal causal models

This Chapter corresponds to the preprint available at https://arxiv.org/abs/1810.01294, and written with Vivian Viallon.

In the Appendix A of the present manuscript, we present preliminary results on natural direct and indirect effects.

Abstract

Many causal models of interest in epidemiology involve longitudinal exposures, confounders and mediators. However, in practice, repeated measurements are not always available. Then, practitioners tend to overlook the time-varying nature of exposures and work under over-simplified causal models. Our objective here was to assess whether - and how - the causal effect identified under such misspecified causal models relates to true causal effects of interest. We focus on two situations regarding the type of available data for exposures: when they correspond to (i) "instantaneous" levels measured at inclusion in the study or (ii) summary measures of their levels up to inclusion in the study. In each of these two situations, we derive sufficient conditions ensuring that the quantities estimated in practice under over-simplified causal models can be expressed as true longitudinal causal effects of interest, or some weighted averages thereof. Unsurprisingly, these sufficient conditions are very restrictive, and our results state that inference based on either "instantaneous" levels or summary measures usually returns quantities that do not directly relate to any causal effect of interest and should be interpreted with caution. They raise the need for repeated measurements and/or the development of sensitivity analyses when such data is not available.

3.1 Introduction

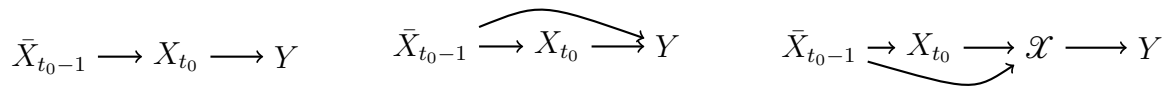
Etiologic epidemiology is concerned with the study of potential causes of chronic diseases based on observational data. Over the years, it has notably been successful in the identification of links between lifestyle exposures and the risk of developing cancer. Remarkable examples are tobacco smoke, alcohol and obesity that are now established risk factors for the development of a number of site-specific cancers (Agudo et al., 2012, Bagnardi et al., 2015, Lauby-Secretan et al., 2016). Moreover, an accumulating body of biomarker measurements and -omics data provide important opportunities for investigating biological mechanisms potentially involved in cancer development. For example, cancer epidemiology is increasingly concerned by the study of the carcinogenic role of inflammation, insulin resistance and sex steroids hormones (Bradbury et al., 2019, Chan et al., 2011, Dossus et al., 2013).

The causal validity of such analyses relies on strong assumptions though, which have been formally described in the causal inference literature (Hernán and Robins, 2020, Pearl, 2000, Robins, 1986, Rosenbaum and Rubin, 1983). The very first assumption underlying most causal analyses is that the causal model is correctly specified. Most often, e.g., when studying lifestyle exposures such as tobacco smoke, alcohol and obesity, but also biomarkers, the true causal model involves time-varying risk factors. Valid causal inference under such longitudinal causal models usually requires repeated measurements for these time-varying variables (Daniel et al., 2012, VanderWeele, 2015, VanderWeele and Tchetgen Tchetgen, 2017). However, such repeated measurements are rarely available in large observational studies, and simplified models that involve time-invariant variables only are usually considered instead. In particular, most studies on biomarkers have been conducted using information collected at recruitment only (Bradbury et al., 2019, Chan et al., 2011, Dossus et al., 2013), since blood samples are usually collected only once, at recruitment, in large cohort studies such as the European Prospective Investigation into Cancer and Nutrition (EPIC) cohort study (Riboli et al., 2002), and the UK Biobank (Sudlow et al., 2015). These studies were conducted after implicitly assuming that past levels of biomarkers are independent of risk of future cancer given current levels of biomarkers; see Figure 3.1 (*L-a*) for a simple illustration, in the absence of confounders. If past levels of biomarkers may influence the outcome not only through their current levels (see, e.g., Figure 3.1 (*L-b*)), the model considered in these analyses was over-simplified, and then misspecified.

Issues arising when working under over-simplified longitudinal causal models have already been described in the statistical literature (Aalen et al., 2016, Maxwell and Cole, 2007, Maxwell et al., 2011). Moreover, general results on the identifiability of causal effects in the presence of unobserved variables can be used to study the identifiability of the causal effect of interest when ignoring the time-varying nature of exposures or, equivalently, when past levels of exposures are unobserved (Huang and Valtorta, 2006, Shpitser and Pearl,

2006, Tian and Pearl, 2002, 2003). However, little is known about the relationship between estimates derived under over-simplified longitudinal causal models and causal quantities of interest under the true longitudinal causal model. Filling this gap is the main objective of the present work. More precisely, we will derive sufficient conditions that guarantee that the quantity estimated in practice when working under misspecified models expresses as a particular weighted average of the longitudinal causal effects of interest. We will consider the most "standard" discrete longitudinal causal models (Daniel et al., 2012), where the causal effect of interest is that of one exposure varying over some predefined discrete time interval, say $[\![1, t_0]\!] := \{1, \ldots, t_0\}$, on one outcome Y measured at some later time point $T > t_0$. Two situations will be considered regarding the available information for the exposures, which will include the exposure of interest and possibly additional factors such as mediators and confounders. First, we consider the situation where available data for the exposures correspond to their "instantaneous" levels at the time t_0 of recruitment in the study. Considering models depicted in Figure 3.1 (*L-a*) and (*L-b*), only data on X_{t_0} would be available, while data on $\bar{X}_{t_0-1}$ would not. This can be regarded as the most common case, but also the worst one since information at one single point in time is available for the full exposure profile. Then, we will turn our attention to a more general and seemingly more favorable situation, where the available information for each exposure corresponds to a summary measure of its levels up to inclusion in the study. Considering exposures such as alcohol intake or dietary exposure, epidemiologists generally not only collect instantaneous levels (through 24-hour recall questionnaires), but also summary measures of past levels of exposure through food frequency questionnaires, which summarize levels of exposures over the last 6 months, 12 months or even 5 years (Slimani et al., 2002). Summary measures are also sometimes constructed from repeated measurements of exposures, when available (Arnold et al., 2016, Kunzmann et al., 2018). This is increasingly common for exposures such as Body Mass Index (BMI) or alcohol intake, whose levels are sometimes available for each participant at different points in time (at recruitment, at 20 years-old, etc.). Cluster analysis can be performed to summarize the repeated measures into a categorical variable, whose categories correspond to certain "shapes" for the exposure profile, such as constantly low, constantly high, etc. Alternatively, the exposure profile can be summarized, e.g., by computing the number of years over a certain threshold, etc. (Arnold et al., 2019, 2016). In any case, the obtained summary measure is then regarded as the exposure of interest, and the underlying time-varying nature of the genuine exposure is not further considered. In other words, these summary measures are supposed to capture everything that matters with respect to the effect of the whole exposure profile on the outcome; see Figure 3.1 (*L-c*) for a simple illustration.

The rest of the article is organized as follows. Section 3.2 presents the notation that will be used throughout the article. In Sections 3.3 and 3.4, we will then present our results, in the situation where instantaneous levels of exposures are available (Section



$(L\text{-}a)$ $(L\text{-}b)$ $(L\text{-}c)$

Figure 3.1: Examples of simple discrete longitudinal causal models with a time-varying exposure $(X_t)_{t\geq 1}$ and an outcome Y, in the absence of confounding. $(L\text{-}a)$ Past levels of exposures $\bar{X}_{t_0-1}$ have no effect on Y, except through current level of exposure X_{t_0}. $(L\text{-}b)$ Past levels of exposures $\bar{X}_{t_0-1}$ have an effect on Y not only through X_{t_0}. $(L\text{-}c)$ The exposure process is assumed to affect the outcome only through some summary variable, $\mathscr{X}$.

3.3), or summary variables of past levels of exposures are available (Section 3.4). We will present concluding remarks and recommendations in Section 3.5. Most technical derivations are presented in the Appendix accompanying this article.

3.2 Notation

For any positive integer i, we use the notation $\mathbf{0}_i$ and $\mathbf{1}_i$ for vectors $(0,\ldots,0) \in \mathbb{R}^i$ and $(1,\ldots,1) \in \mathbb{R}^i$ respectively. As mentioned above, we consider the setting that is classically adopted when working with time-varying predictors in causal inference (Daniel et al., 2012, VanderWeele, 2015). More precisely, we assume that time-varying exposures, including the exposure of interest as well as potential mediators and confounders, are observable at discrete times over the time-window $[\![1;T]\!] := \{1,\ldots,T\}$ for some $T > 1$. For any $t \in [\![1;T]\!]$, we let X_t denote the exposure of interest at time t. Adopting the notation of VanderWeele (2015), we further denote the exposure profile until time t by $\bar{X}_t = (X_1, X_1, \ldots, X_t)$, while $\bar{x}_t$ stands for a specific (fixed) profile for the exposure of interest. Full exposure profile is denoted by $\bar{X} = \bar{X}_T = (X_1, X_2, \ldots X_T)$. When needed, we will use similar notation for auxiliary factors $(Z_t)_{t\geq 1}$, that may include pure mediator processes $(M_t)_{t\geq 1}$, as well as confounder processes $(W_t)_{t\geq 1}$ possibly affected by the exposure of interest. Unless otherwise stated, we assume that all the variables are binary to simplify the notation. We further denote by $t_0 \in [\![2;T]\!]$ the inclusion time in the study.

While causal inference should generally rely on the observations of the full profile of exposures $(\bar{X}, \bar{Z})$, or at least their full profile prior to inclusion $(\bar{X}_{t_0}, \bar{Z}_{t_0})$, we assume in Section 3.3 that the available information at time t_0 consists in (X_{t_0}, Z_{t_0}) only. Next, Section 3.4 will be devoted to the case where we have access to some summary measures of $\bar{X}_{t_0}$ and $\bar{Z}_{t_0}$, which will be denoted by $\mathscr{X}$ and $\mathscr{Z}$, respectively. These summary measures are typically defined as deterministic functions of the exposure profiles. Considering, e.g., summary measures of $\bar{X}_{t_0}$, typical examples include functions of the form $\mathscr{X} = \sum_{t=t'}^{t_0} X_t$, and $\mathscr{X} = \mathbb{1}\{\sum_{t=t'}^{t_0} X_t \geq \tau\}$ for some $1 \leq t' \leq t_0$ and some threshold $\tau \in \mathbb{R}$. More simply,

we can even have $\mathscr{X} = X_{t_0}$, which emphasizes the fact that the situation where summary measures are available encompasses the situation where instantaneous levels are available as a special case.

For any pair of variables (V, U) and any potential value u of U, we denote by $V^{U=u}$ the counterfactual variable corresponding to variable V that would have been observed in the counterfactual world following the hypothetical intervention $do(U = u)$. We work under the setting of Structural Causal Models (Pearl, 1995, 2000), which especially entails that consistency conditions hold: for instance, $U = u$ implies $V = V^{U=u}$. In addition, we assume that positivity conditions hold (Rosenbaum and Rubin, 1983). For any possibly counterfactual random variables V and U, and any causal model (Mod), we will use the notation $(V \perp\!\!\!\perp U)_{Mod}$ to denote independence between variables V and U under the causal model (Mod). We will further let $\mathbb{E}_{Mod}\left(V^{U=u}\right)$ be the expectation of variable $V^{U=u}$ under causal model (Mod). We will mostly consider such expectations for Mod set to either the true causal longitudinal model (we will use indices L and LS for these longitudinal models when considering models involving instantaneous levels only, and summary variables, respectively) or the over-simplified model used for the analysis (we will use indices CS - standing for cross-sectional - and SV - standing for summary variables - for these models). In particular, a key quantity in our work is $ATE_L\left(\bar{x}_t; \bar{x}_t^*\right) = \mathbb{E}_L(Y^{\bar{X}_t=\bar{x}_t} - Y^{\bar{X}_t=\bar{x}_t^*})$, for any two given profiles $\bar{x}_t$ and $\bar{x}_t^*$ for the exposure of interest, and some time t. This quantity is one measure of the total effect (Daniel et al., 2012, VanderWeele, 2015) of exposure up to time t on the outcome variable Y under a given longitudinal causal model (L), as for instance the one given in Figure 3.1 $(L\text{-}b)$. More details will be be given in Section 3.3. Because this quantity generally depends on the particular values for $\bar{x}_t$ and $\bar{x}_t^*$, averaged total effects can be defined for appropriate weights $\omega(\bar{x}_t, \bar{x}_t^*)$ as $\sum_{\bar{x}_t}\sum_{\bar{x}_t^*} ATE_L\left(\bar{x}_t; \bar{x}_t^*\right)\omega(\bar{x}_t, \bar{x}_t^*)$, with the two sums over $\{0,1\}^t$. We will also consider stratum-specific causal effects (Hernán and Robins, 2020), with strata defined according to the levels of some possibly multivariate variable U

$$ATE_{L_{|U=u}}\left(\bar{x}_t; \bar{x}_t^*\right) := \mathbb{E}_L\left(Y^{\bar{X}_t=\bar{x}_t} - Y^{\bar{X}_t=\bar{x}_t^*} \mid U = u\right), \tag{3.1}$$

and weighted averages of the form $\sum_u \sum_{\bar{x}_t}\sum_{\bar{x}_t^*} ATE_{L_{|U=u}}\left(\bar{x}_t; \bar{x}_t^*\right)\omega(\bar{x}_t, \bar{x}_t^*, u)$, for appropriate weights $\omega(\bar{x}_t, \bar{x}_t^*, u)$.

Then, we need to introduce a specific symbol, $\fallingdotseq$, to relate a causal effect defined under some over-simplified model to the quantity that is actually estimated in practice, and which is usually expressed under the true longitudinal causal model. Consider, e.g., an over-simplified causal model (CS), under which the causal effect $ATE_{CS} := \mathbb{E}_{CS}\left(Y^{X_{t_0}=1} - Y^{X_{t_0}=0}\right)$ can be identified through the formula $\mathbb{E}_{CS}\left(Y \mid X_{t_0} = 1\right) - \mathbb{E}_{CS}\big(Y \mid X_{t_0} = 0\big)$. Because this quantity will actually be estimated using data generated under the true longitudinal model, say (L), the quantity estimated in practice turns out to be $\mathbb{E}_L\left(Y \mid X_{t_0} = 1\right) - \mathbb{E}_L\left(Y \mid X_{t_0} = 0\right)$. We would then write $ATE_{CS} \fallingdotseq \mathbb{E}_L\left(Y \mid X_{t_0} = 1\right) -$

$\mathbb{E}_L(Y \mid X_{t_0}=0)$. We shall stress that $ATE_{CS} \fallingdotseq \mathbb{E}_L(Y \mid X_{t_0}=1) - \mathbb{E}_L(Y \mid X_{t_0}=0)$ does generally not imply $ATE_{CS} = \mathbb{E}_L(Y \mid X_{t_0}=1) - \mathbb{E}_L(Y \mid X_{t_0}=0)$, unless, e.g., (CS) is correctly specified. For the sake of legibility, we will indistinctly use ATE_{CS} for both the causal effect and the quantity estimated in practice in the text.

In other respect, expectations and probabilities involving observed variables only will from now on be computed under the true longitudinal causal model, and so we will simply use notation like $\mathbb{E}(V)$ and $\mathbb{P}(V=v)$ for any observable variable V. Going back to the example above, we would therefore simply write $ATE_{CS} \fallingdotseq \mathbb{E}(Y \mid X_{t_0}=1) - \mathbb{E}(Y \mid X_{t_0}=0)$, which means that the quantity estimated in practice when working under the over-simplified causal model (CS) is actually $\mathbb{E}_L(Y \mid X_{t_0}=1) - \mathbb{E}_L(Y \mid X_{t_0}=0)$. See, e.g., the proof of Theorem 1 in Appendix 3.A.1 for more details.

Finally, in our causal diagrams, we will use as usual simple solid arrows $U \longrightarrow V$ to denote that U is a potential cause of V, for any possibly multivariate random variables U and V. In addition, double dashed arrows $V \dashleftarrow\!\!\dashrightarrow U$ will be used when (i) components of U may cause components of V, (ii) components of U may be caused by components of V, but (iii) any univariate component $\tilde{U} \subset U$ causing a univariate component $\tilde{V} \subset V$ cannot be caused by $\tilde{V}$. See Figure 3.2 (L) for a simple example of a causal diagram involving such double dashed arrows. We shall stress that our double dashed arrows have a different meaning than the usual dashed double-headed arrow $V \dashleftarrow\!\dashrightarrow U$ used in the literature (Shpitser and Pearl, 2006, Tian and Pearl, 2002, 2003) when the $(U-V)$ relationship may be confounded by unmeasured variables. Moreover, point (iii) ensures that the subgraph $V \dashleftarrow\!\!\dashrightarrow U$ is still a directed acyclic graph (DAG).

3.3 The case when exposure variables are measured at inclusion in the study only

3.3.1 General model and results

A general causal model where a time-varying exposure $(X_t)_{t\geq 1}$ potentially causes an outcome Y, can be compactly represented as in Figure 3.2 (L). Here, variables $(Z_t)_{t\geq 1}$ are possibly multivariate, in which case their components may consist of pure mediators, pure confounders, as well as confounders influenced by the exposure of interest. Moreover, some components of Z_{t_0} may be unobserved in practice. At each time $t \in [\![1;T]\!]$, X_t is a potential cause of Y and is potentially caused by all components of $\bar{X}_{t-1}$, and by some or all components of $\bar{Z}_{t-1}$ and Z_t. At each time $t \in [\![1;T]\!]$, Z_t is a potential cause of Y, whose components are potentially caused by $\bar{X}_{t-1}$ and $\bar{Z}_{t-1}$. Components of Z_t that are not causes of X_t may further be caused by X_t. This general model could depict the case where the exposure of interest $(X_t)_{t\geq 1}$ stands for BMI at different ages, while the auxiliary variable $(Z_t)_{t\geq 1}$ would include measures of alcohol intake, physical activity and

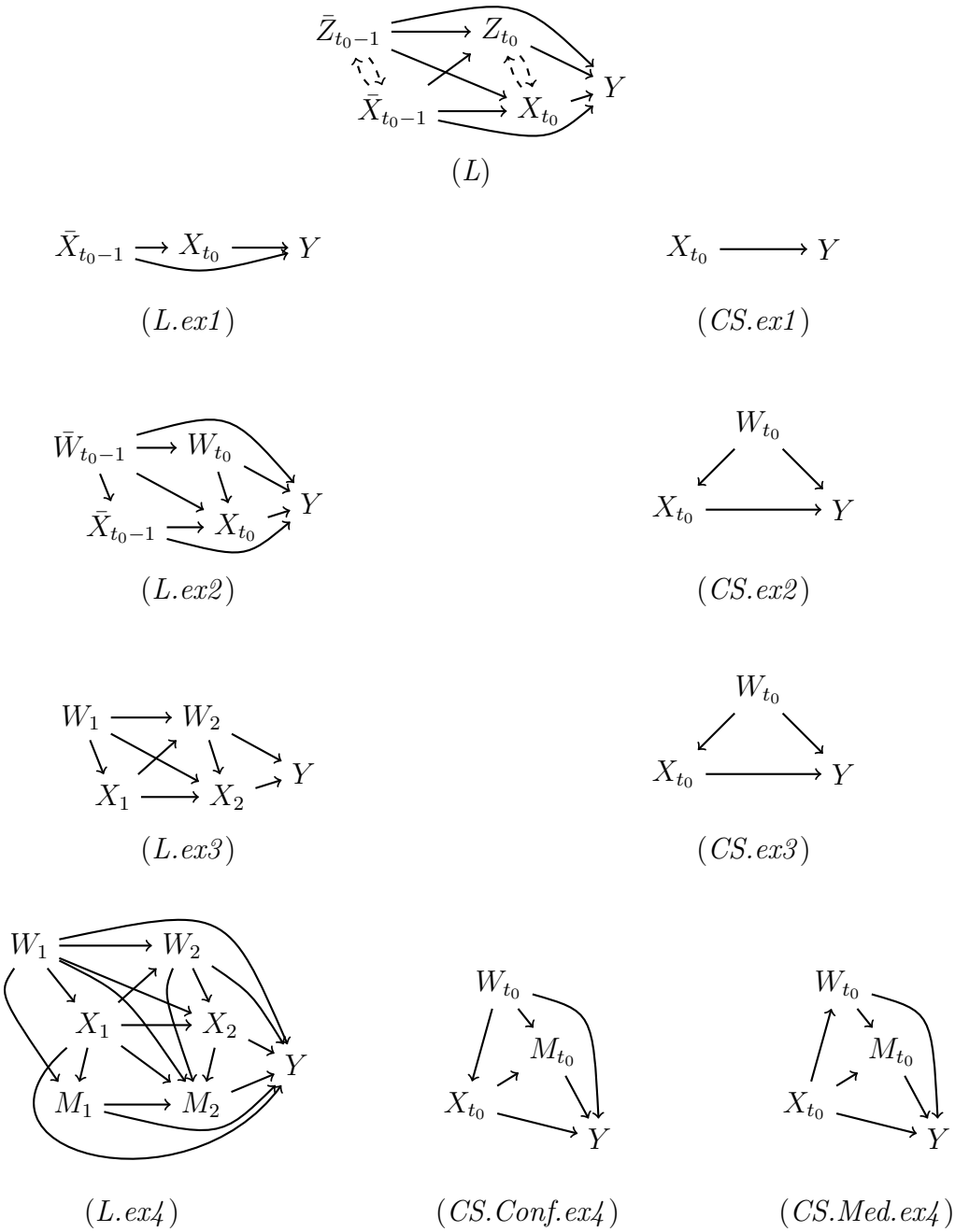


Figure 3.2: (L) General longitudinal causal model with time-varying exposure of interest $(X_t)_{t\geq 1}$, and additional time-varying process $(Z_t)_{t\geq 1}$. Particular cases are presented in $(L.ex1)$, $(L.ex2)$, $(L.ex3)$ and $(L.ex4)$, along with their over-simplified counterparts in $(CS.Conf.ex1)$, $(CS.Conf.ex2)$, $(CS.Conf.ex3)$, $(CS.Conf.ex4)$, and $(CS.Med.ex4)$. When the true longitudinal model is $(L.ex4)$, with a time-varying confounder $(W_t)_{t\geq 1}$ affected by the exposure, two possible over-simplified counterparts can be considered, depending on whether $(W_t)_{t\geq 1}$ is mainly considered as a confounder or a mediator.

diet at different ages. Model ($L.ex4$) in Figure 3.2 provides a less compact representation of a particular example of this general model, with $t_0 = 2$, and $Z_t = (M_t, W_t)$, where $(W_t)_{t\geq 1}$ is a confounder affected by the exposure, and $(M_t)_{t\geq 1}$ a pure mediator.

Under such models, causal effects can be defined by considering hypothetical interventions on the full exposure profile $do(\bar{X} = \bar{x})$. However, epidemiologists are often interested in the assessment of the predictive role of the exposure of interest, so a more natural measure of the causal effect of exposure on the outcome is

$$ATE_L\left(\bar{x}_{t_0}; \bar{x}^*_{t_0}\right) := \mathbb{E}_L\left(Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} - Y^{\bar{X}_{t_0}=\bar{x}^*_{t_0}}\right), \tag{3.2}$$

for any given exposure profiles up to time t_0, $\bar{x}_{t_0}$ and $\bar{x}^*_{t_0}$ in $\{0,1\}^{t_0}$. Under some well known sets of assumptions on the causal model, including the consistency and sequential ignorability conditions (Pearl, 2000, Robins, 1986, Rosenbaum and Rubin, 1983), the causal effect in Equation (3.2) can be expressed in terms of observable variables only. It can then be estimated if data on the full history of the variables up to time t_0 is available, assuming that some positivity conditions hold (Rosenbaum and Rubin, 1983). We recall that such positivity conditions will be assumed to hold throughout this article.

However, when data on exposures are available at time t_0 only, $ATE_L\left(\bar{x}_{t_0}; \bar{x}^*_{t_0}\right)$ can generally not be estimated. As mentioned in the Introduction 3.1, it is then common practice to implicitly (i) overlook the time-varying nature of the exposures, (ii) work under an over-simplified causal model (CS), and (iii) consider the causal effect $ATE_{CS} := \mathbb{E}_{CS}\left(Y^{X_{t_0}=1} - Y^{X_{t_0}=0}\right)$ as the causal measure of interest. For example, if the true causal longitudinal model is model ($L.ex4$) of Figure 3.2, but only information on Y, X_{t_0}, M_{t_0} and W_{t_0} is available, most practitioners would implicitly work under the over-simplified model ($CS.Conf.ex4$) given in Figure 3.2. Then, because $(Y^{X_{t_0}=x_{t_0}} \perp\!\!\!\perp X_{t_0}|W_{t_0})_{CS.Conf.ex4}$, the quantity of interest would be identified through

$$ATE_{CS.Conf.ex4} \;\fallingdotseq\; \sum_{w_{t_0}} \left[\mathbb{E}\left(Y \mid W_{t_0} = w_{t_0}, X_{t_0} = 1\right) - \mathbb{E}\left(Y \mid W_{t_0} = w_{t_0}, X_{t_0} = 0\right)\right] \times \mathbb{P}(W_{t_0} = w_{t_0}).$$

It is noteworthy that, for some true longitudinal causal models, several over-simplified cross-sectional models may be considered. When the true causal longitudinal model is that of Figure 3.2 ($L.ex4$), practitioners may consider $(W_t)_t$ mainly as a confounder and work with the over-simplified model ($CS.Conf.ex4$), but they may also consider $(W_t)_t$ mainly as a mediator and work with model ($CS.Med.ex4$). Because $(Y^{X_{t_0}=x_{t_0}} \perp\!\!\!\perp X_{t_0})_{CS.Med.ex4}$, the quantity estimated in practice in the latter case would be $ATE_{CS.Med.ex4} \fallingdotseq \mathbb{E}\left(Y \mid X_{t_0} = 1\right) - \mathbb{E}\left(Y \mid X_{t_0} = 0\right)$.

Then, a natural question is whether - and how - the quantity estimated in practice when working under over-simplified causal models (CS) relates to the longitudinal causal effects under the true model (L), or to another causal effect of interest. Theorem 1 below

presents a sufficient condition under which the quantity estimated in practice actually equals $ATE_L(1;0) := \mathbb{E}_L\left(Y^{X_{t_0}=1} - Y^{X_{t_0}=0}\right)$, the causal effect of X_{t_0} (which of course usually differs from that of $\bar{X}_{t_0}$, but can still be seen as a causal effect of interest). Theorem 2 then presents a weaker sufficient condition under which ATE_{CS} expresses as a weighted average of stratum specific longitudinal total effects (3.1). Detailed proofs of these results are given in Appendix 3.A.1 for Theorem 1, and Appendix 3.A.2 for Theorem 2. In Section 3.3.2 below, we illustrate their implications by focusing on a few simple examples.

Theorem 1. *If condition* $(T1.Cond)$ *below holds*

$(T1.Cond)$ *There exists some observed* $W_{t_0} \subset Z_{t_0}$ *taking values in* $\Omega_{W_{t_0}}$*, such that* $(Y^{X_{t_0}=x_{t_0}} \perp\!\!\!\perp X_{t_0}|W_{t_0})_{CS}$ *and* $(Y^{X_{t_0}=x_{t_0}} \perp\!\!\!\perp X_{t_0}|W_{t_0})_L$

then the quantity estimated in practice equals $ATE_L(1;0) = \mathbb{E}_L\left(Y^{X_{t_0}=1} - Y^{X_{t_0}=0}\right)$*:*

$$\begin{aligned} ATE_{CS} &\fallingdotseq \sum_{w_{t_0}\in\Omega_{W_{t_0}}} [\mathbb{E}(Y \mid W_{t_0}=w_{t_0}, X_{t_0}=1) - \mathbb{E}(Y \mid W_{t_0}=w_{t_0}, X_{t_0}=0)] \\ &\quad \times \mathbb{P}(W_{t_0}=w_{t_0}), \qquad (3.3) \\ &= ATE_L(1;0). \qquad (3.4) \end{aligned}$$

In particular, if condition $(T1.Uncond)$ *below holds*

$(T1.Uncond)$ $(Y^{X_{t_0}=x_{t_0}} \perp\!\!\!\perp X_{t_0})_{CS}$ *and* $(Y^{X_{t_0}=x_{t_0}} \perp\!\!\!\perp X_{t_0})_L$

then

$$ATE_{CS} \fallingdotseq \mathbb{E}(Y \mid X_{t_0}=1) - \mathbb{E}(Y \mid X_{t_0}=0) = ATE_L(1;0).$$

Theorem 2. *If condition* $(T2.Cond)$ *below holds*

$(T2.Cond)$ *There exists some observed* $W_{t_0} \subset Z_{t_0}$ *taking values in* $\Omega_{W_{t_0}}$*, such that* $(Y^{X_{t_0}=x_{t_0}} \perp\!\!\!\perp X_{t_0}|W_{t_0})_{CS}$ *and* $(Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} \perp\!\!\!\perp \bar{X}_{t_0} \mid W_{t_0})_L$

then the quantity estimated in practice

$$\begin{aligned} ATE_{CS} &\fallingdotseq \sum_{w_{t_0}\in\Omega_{W_{t_0}}} [\mathbb{E}(Y \mid W_{t_0}=w_{t_0}, X_{t_0}=1) - \mathbb{E}(Y \mid W_{t_0}=w_{t_0}, X_{t_0}=0)] \\ &\quad \times \mathbb{P}(W_{t_0}=w_{t_0}), \\ &= \sum_{w_{t_0}\in\Omega_{W_{t_0}}} \sum_{\substack{\bar{x}_{t_0-1}\in\{0,1\}^{t_0-1} \\ \bar{x}^*_{t_0-1}\in\{0,1\}^{t_0-1}}} \Big\{ ATE_{L_{|W_{t_0}=w_{t_0}}}\left((\bar{x}_{t_0-1},1);(\bar{x}^*_{t_0-1},0)\right) \\ &\qquad \times \mathbb{P}(\bar{X}_{t_0-1}=\bar{x}_{t_0-1} \mid X_{t_0}=1, W_{t_0}=w_{t_0}) \\ &\qquad \times \mathbb{P}(\bar{X}_{t_0-1}=\bar{x}^*_{t_0-1} \mid X_{t_0}=0, W_{t_0}=w_{t_0}) \\ &\qquad \times \mathbb{P}(W_{t_0}=w_{t_0})\Big\}. \qquad (3.5) \end{aligned}$$

In particular, if condition $(T2.Uncond)$ *below holds*

$(T2.Uncond)$ $\quad (Y^{X_{t_0}=x_{t_0}} \perp\!\!\!\perp X_{t_0})_{CS}$ *and* $(Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} \perp\!\!\!\perp \bar{X}_{t_0})_{L}$

then

$$\begin{aligned} ATE_{CS} &\fallingdotseq \mathbb{E}\left(Y \mid X_{t_0}=1\right) - \mathbb{E}\left(Y \mid X_{t_0}=0\right) \\ &= \sum_{\substack{\bar{x}_{t_0-1}\in\{0,1\}^{t_0-1} \\ \bar{x}^*_{t_0-1}\in\{0,1\}^{t_0-1}}} \begin{aligned}[t] &\big\{ATE_L\left((\bar{x}_{t_0-1},1);(\bar{x}^*_{t_0-1},0)\right) \\ &\times \mathbb{P}(\bar{X}_{t_0-1}=\bar{x}_{t_0-1} \mid X_{t_0}=1) \\ &\times \mathbb{P}(\bar{X}_{t_0-1}=\bar{x}^*_{t_0-1} \mid X_{t_0}=0)\big\}. \end{aligned} \end{aligned} \tag{3.6}$$

Theorem 1 states that whenever there exists a set of observed variables that satisfies the ignorability condition for the exposure at time t_0, X_{t_0}, and the outcome under both the true and over-simplified causal models, then, the quantity estimated in practice equals the longitudinal total effect $ATE_L\left(1;0\right)$. In the same way, Theorem 2 states that whenever there exists a set of observed variables that satisfies (i) the ignorability condition for the whole time-varying exposure profile, $\bar{X}_{t_0}$, and the outcome under the true longitudinal model, and (ii) the ignorability condition for the exposure at time t_0, X_{t_0}, and the outcome under the over-simplified causal model, then the quantity estimated in practice can be written in terms of stratum specific longitudinal total effects.

3.3.2 Examples and illustration of the general results

When the conditions of Theorem 1 and Theorem 2 are not satisfied, the quantity estimated in practice has to be interpreted with caution as its relationship with causal effects of interest usually remains unclear. See for example Web Supplementary Material 3.C.1 where the case of the model $(L.ex2)$ given in Figure 3.2 is described in details. However, the conditions of our Theorems being sufficient conditions only, there are a few cases where they are not satisfied but ATE_{CS} is still an informative measure of the exposure effect. For example, denote by $(L.ex2')$ and $(CS.ex2')$ the versions of $(L.ex2)$ and $(CS.ex2)$, respectively, after removing the arrow from X_{t_0} to Y. In this particular case where X_{t_0} has no causal effect on Y and only a pure time-varying confounder is present, we have $(Y \perp\!\!\!\perp \bar{X}_{t_0} \mid \bar{W}_{t_0})_{L.ex2'}$, but we do not have $(Y \perp\!\!\!\perp \bar{X}_{t_0} \mid W_{t_0})_{L.ex2'}$, so the conditions of our Theorems are not satisfied. Nevertheless, we still have $ATE_{CS.ex2'}=0$, and the inference under the over-simplified model is valid. However, we shall stress that ATE_{CS} can also be null in other situations where the exposure does affect the outcome, even when the condition of Theorem 2 is satisfied (we will get back to this point below).

When the conditions of Theorem 1 are satisfied, the interpretation of ATE_{CS} is straightforward as it simply equals $ATE_L(1;0)$. However, unsurprisingly, these conditions are very restrictive. For example, the condition $(Y^{X_{t_0}=x_{t_0}} \perp\!\!\!\perp X_{t_0} \mid W_{t_0})_L$ is generally not satisfied under models $(L.ex1)$, $(L.ex2)$ and $(L.ex4)$ given in Figure 3.2, because $\bar{X}_{t_0-1}$, and possibly $\bar{W}_{t_0-1}$, act as unmeasured confounders for the $(X_{t_0}-Y)$ relationship, but

are ignored in the over-simplified models. On the other hand, the conditions of Theorem 1 are verified under the very simple models $(L-a)$ of Figure 3.1 and $(L.ex3)$ of Figure 3.2, as well as under particular cases of model $(L.ex2)$, e.g., when there is no arrow from $\bar{X}_{t_0-1}$ to Y nor from $\bar{W}_{t_0-1}$ to Y.

Before discussing the interpretation of ATE_{CS} under the conditions of Theorem 2, we shall stress that these conditions are quite restrictive too. In particular, they are not satisfied under model $(L.ex4)$ of Figure 3.2, where $(W_t)_{t>1}$ is a confounder affected by the exposure. Under this model, sequential ignorability holds: $(Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} \perp\!\!\!\perp X_1 \mid W_1)_{L.ex4}$ and $\left(Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} \perp\!\!\!\perp \bar{X}_t \mid \{\bar{W}_t, \bar{X}_{t-1}\}\right)_{L.ex4}$ for any $t \in [\![2; t_0]\!]$ (Daniel et al., 2012, Hernán and Robins, 2020, Robins, 1986). But the conditions of Theorem 2 are not satisfied: we neither have $(Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} \perp\!\!\!\perp \bar{X}_{t_0} \mid \bar{W}_{t_0})_{L.ex4}$ nor $(Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} \perp\!\!\!\perp \bar{X}_{t_0} \mid W_{t_0})_{L.ex4}$, because $(W_t)_{t>1}$ acts as both a confounder and a mediator in the $(\bar{X}_{t_0} - Y)$ relationship. The conditions of Theorem 2 are generally not satisfied either under model $(L.ex2)$ of Figure 3.2, because $\bar{W}_{t_0-1}$ affects Y not through W_{t_0} and X_{t_0} only, and therefore acts as an unmeasured confounder in the $(X_{t_0} - Y)$ relationship, which is ignored in model $(CS.ex2)$.

We will now discuss the interpretability of ATE_{CS} when conditions of Theorem 2 are satisfied by focusing on the simple example of model $(L.ex1)$ and its simplified counterpart $(CS.ex1)$ in Figure 3.2. Here, we have $(Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} \perp\!\!\!\perp \bar{X}_{t_0})_{L.ex1}$ and $(Y^{X_{t_0}=x_{t_0}} \perp\!\!\!\perp X_{t_0})_{CS.ex1}$, so that Theorem 2 ensures that $ATE_{CS.ex1}$ is a weighted sum of the longitudinal total effects that compare any possible pairs of exposure profiles up to time t_0, one of which terminating with $X_{t_0} = 0$ and the other one with $X_{t_0} = 1$ (see Equation (3.6)). However, the relevance of this particular weighted average is generally questionable. Indeed, because of the non-negative weights for terms like $ATE_{L.ex1}\left((\mathbf{0}_{t_0-1}, 1); (\mathbf{1}_{t_0-1}, 0)\right)$, $ATE_{CS.ex1}$ can be null even for models under which each X_t, for $t = 1, \ldots, t_0$, has a, say, positive effect on Y. This particular case illustrates that ATE_{CS} generally has to be interpreted with caution even when conditions of Theorem 2 are satisfied.

The interpretation of the weighted average in Equation (3.6) is more straightforward if profiles $\bar{x}_{t_0-1}$ associated with large weights $\mathbb{P}(\bar{X}_{t_0-1} = \bar{x}_{t_0-1} \mid X_{t_0} = 1)$ correspond to globally more exposed profiles than the profiles $\bar{x}^*_{t_0-1}$ associated with large weights $\mathbb{P}(\bar{X}_{t_0-1} = \bar{x}^*_{t_0-1} \mid X_{t_0} = 0)$. In particular, this is the case when the exposure is "stable", more precisely when $X_t = 1 \Rightarrow X_{t'} = 1$ for all $t' \geq t$. This stability assumption can be seen as a reasonable assumption (or approximation) for exposures such as obesity for instance. When it is satisfied, the only exposure profile that terminates with $x_{t_0} = 0$ is $\bar{x}_{t_0} = \mathbf{0}_{t_0}$, and, under model $(L.ex1)$, ATE_{CS} then reduces to

$$\sum_{i=0}^{t_0-1} ATE_L\left((\mathbf{0}_i, \mathbf{1}_{t_0-i}); \mathbf{0}_{t_0}\right) \times \mathbb{P}\left(\bar{X}_{t_0-1} = (\mathbf{0}_i, \mathbf{1}_{t_0-i-1}) \mid X_{t_0} = 1\right). \tag{3.7}$$

The stability assumption then guarantees that ATE_{CS} is a weighted sum of all the longitudinal causal effects comparing the ever-exposed profiles to the single never-exposed

profile. Weights in the equation above are sensible as they correspond to the actual proportions of subjects with exposure profiles $(\mathbf{0}_i, \mathbf{1}_{t_0-i})_{i\in[\![0,t_0-1]\!]}$ among the subpopulation of exposed individuals at time t_0. Therefore, ATE_{CS} can be regarded as a meaningful quantity under model (*L.ex1*) if the stability assumption further holds. The fact that ATE_{CS} is a meaningful quantity under the stability assumption extends to the situation where a time-invariant observed confounder W is added to model (*L.ex1*). However, we recall that if the confounder is time-varying, as in Figure 3.2 (*L.ex*2), the conditions of Theorem 2 are not satisfied, and ATE_{CS} has usually no clear meaning, even when both the exposure and confounder processes are stable. We refer to Web Supplementary Material 3.C.1 for more details on this particular case.

To recap, when only instantaneous levels of exposures at inclusion are available, the quantity estimated in practice when working under over-simplified models has generally to be interpreted with caution, even when the conditions of Theorem 2 are satisfied. Except for a few exceptions, and unsurprisingly, the quantity estimated in practice can only be unambiguously related to causal effects of interest when the conditions of Theorem 1 are satisfied. We have shown this was notably the case under model (*L-a*) of Figure 3.1, where the effect of $\bar{X}_{t_0}$ on the outcome is entirely mediated by X_{t_0}. Interestingly, this situation arises as a particular case of the model presented in Figure 3.1 (*L-c*) where a summary variable $\mathscr{X}$ is assumed to mediate the whole effect of $\bar{X}_{t_0}$ on the outcome. In the following Section, we consider more general situations where data collected at time t_0 corresponds to such summary measures of past levels of exposures, as is sometimes assumed, or implicitly assumed, in epidemiological studies.

3.4 The case when summaries of past levels of exposures are available

3.4.1 General models and results

We will now turn our attention to the situation where data collected at time t_0 concerns summary measures of past levels of exposures, and where the whole effect of exposures on the outcome Y is captured by these summary measures (Arnold et al., 2019, 2016, De Rubeis et al., 2019, Fan et al., 2008, Kunzmann et al., 2018, Platt et al., 2010, Yang et al., 2019, Zheng et al., 2018). A general representation of such models is given in Figure 3.3 (*LS*), where, as in the previous Section, $(Z_t)_{t\geq 1}$ can be multivariate, and so can $\mathscr{Z}$. Moreover, some components of $\mathscr{Z}$ may be unobserved. Again, $(X_t)_{t\geq 1}$ could stand for BMI at different ages, and $(Z_t)_{t\geq 1}$ could include measures of alcohol intake, physical activity and diet at different ages, while $\mathscr{X}$ and $\mathscr{Z}$ would be any appropriate summary measures of $\bar{X}_{t_0}$ and $\bar{Z}_{t_0}$, respectively. The simplest model of this form is the one given in Figure 3.1 (*L-c*), and corresponds to the absence of any confounding process. Other

examples are given in Figure 3.3; we will present them in more details below.

Let us first discuss the causal effects of interest in this setting. Distinct exposure profiles $\bar{x}_{t_0}$ leading to $\mathscr{X} = x$, for any potential value x of $\mathscr{X}$, can be seen as distinct versions of the "compound treatment" x (Hernán and VanderWeele, 2011, VanderWeele and Hernán, 2013), in the particular case where versions precede what we will refer to as treatment $\mathscr{X}$, or x, below. Moreover, because summary variables are deterministic functions of exposure profiles, interventions on the latter, but not on the former, can be implemented in practice. As a result, $Y^{\mathscr{X}=x}$, although mathematically grounded, may not have a clear practical meaning. Then, and as we will now describe, causal effects of natural interest under models involving summary variables actually depend on whether or not the versions of $\mathscr{X}$ are relevant (Hernán and VanderWeele, 2011).

Adopting the same terminology as Hernán and VanderWeele (2011), we will say that versions of treatment $\mathscr{X}$ are irrelevant, when all versions $\bar{x}_{t_0}$ leading to $\mathscr{X} = x$ also lead to the same effect on the outcome, or, more precisely when condition $(Irrel)$ below holds:

$$(Irrel) \quad Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} = Y^{\mathscr{X}=x} \text{ for any } \bar{x}_{t_0} \text{ such that } \bar{X}_{t_0} = \bar{x}_{t_0} \Rightarrow \mathscr{X} = x.$$

When the versions are irrelevant, as in model $(L\text{-}c)$ of Figure 3.1 for example, we have $ATE_{LS}(\bar{x}_{t_0}; \bar{x}^*_{t_0}) = ATE_{LS}(x; x^*) = \mathbb{E}_{LS}\big(Y^{\mathscr{X}=x} - Y^{\mathscr{X}=x^*}\big)$, for any $\bar{x}_{t_0}$ and $\bar{x}^*_{t_0}$ leading to $\mathscr{X} = x$ and $\mathscr{X} = x^*$, respectively. As a result, $\mathbb{E}_{LS}\left(Y^{\mathscr{X}=x} - Y^{\mathscr{X}=x^*}\right)$ is well-defined and constitutes a causal effect of interest.

On the other hand, we will say that versions of the treatment are relevant when $Y^{\bar{X}_{t_0}=\bar{x}_{t_0}}$ and $Y^{\bar{X}_{t_0}=\bar{x}'_{t_0}}$ may be different even though $\bar{x}_{t_0}$ and $\bar{x}'_{t_0}$ are two exposure profiles leading to the same value x for $\mathscr{X}$. For example, this is typically the case under model (LS) of Figure 3.3, since $\bar{X}_{t_0}$ affects Y not only through $\mathscr{X}$ but also through some components of $\mathscr{Z}$. Indeed, we can have $\mathscr{Z}^{\bar{X}_{t_0}=\bar{x}_{t_0}} \neq \mathscr{Z}^{\bar{X}_{t_0}=\bar{x}'_{t_0}}$, and, in turn $Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} \neq Y^{\bar{X}_{t_0}=\bar{x}'_{t_0}}$, for some exposure profiles $\bar{x}_{t_0}$ and $\bar{x}'_{t_0}$ leading to the same value $\mathscr{X} = x$. Then, when versions are relevant, we typically have $ATE_{LS}(\bar{x}_{t_0}; \bar{x}^*_{t_0}) \neq ATE_{LS}(\bar{x}'_{t_0}; \bar{x}'^*_{t_0})$, even if both $\bar{x}_{t_0}$ and $\bar{x}'_{t_0}$ lead to $\mathscr{X} = x$ and both $\bar{x}^*_{t_0}$ and $\bar{x}'^*_{t_0}$ lead to $\mathscr{X} = x^*$. Therefore, although still mathematically grounded, the quantity $ATE_{LS}(x; x^*) = \mathbb{E}_{LS}\left(Y^{\mathscr{X}=x} - Y^{\mathscr{X}=x^*}\right)$ is not well defined from a "practical point of view", and cannot be considered as a quantity of interest. Among other possibilities, the quantity

$$\begin{aligned}\sum_{x_{t_0}} \{\mathbb{E}_{LS}\big(Y^{\bar{x}_{t_0}}\big) \times \mathbb{P}(\bar{X}_{t_0} = \bar{x}_{t_0} \mid \mathscr{X} = x)\} - \sum_{x^*_{t_0}} \{\mathbb{E}_{LS}\big(Y^{\bar{x}^*_{t_0}}\big) \times \mathbb{P}(\bar{X}_{t_0} = \bar{x}^*_{t_0} \mid \mathscr{X} = x^*)\} \\ = \sum_{\bar{x}_{t_0}} \sum_{\bar{x}^*_{t_0}} \{ATE_{LS}(\bar{x}_{t_0}; \bar{x}^*_{t_0}) \times \mathbb{P}(\bar{X}_{t_0} = \bar{x}_{t_0} \mid \mathscr{X} = x) \times \mathbb{P}(\bar{X}_{t_0} = \bar{x}^*_{t_0} \mid \mathscr{X} = x^*)\}, \quad (3.8)\end{aligned}$$

can be regarded as a causal effect of interest. It corresponds to the difference between the expectation of the outcome in the following two counterfactual populations. In the first one, for any profile $\bar{x}_{t_0}$ leading to $\mathscr{X} = x$, a proportion $\mathbb{P}(\bar{X}_{t_0} = \bar{x}_{t_0} \mid \mathscr{X} = x)$ of the individuals undergo the intervention $do(\bar{X}_{t_0} = \bar{x}_{t_0})$. This can be regarded as a natural way

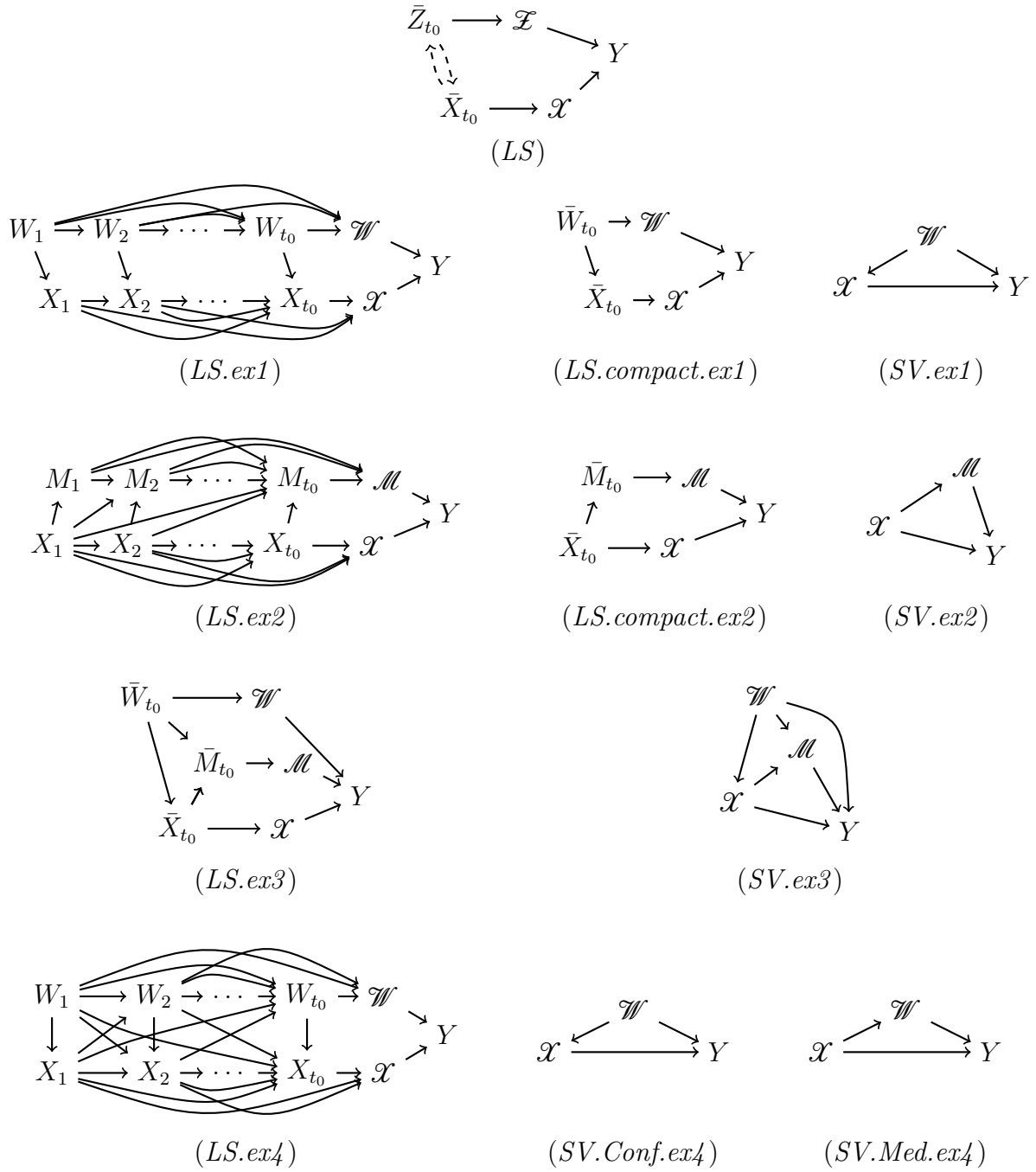


Figure 3.3: (LS) General longitudinal causal model with time-varying exposure of interest $(X_t)_{t\geq 1}$, and a additional time-varying process $(Z_t)_{t\geq 1}$, in the situation where exposure profiles only affect the outcome through some summary variables $\mathscr{X}$ and $\mathscr{Z}$. Particular cases are presented in $(LS.ex1)$ (or more compactly in $(LS.compact.ex1)$), , $(LS.ex2)$ (or more compactly in $(LS.compact.ex2)$), $(LS.ex3)$ and $(LS.ex3)$, along with their over-simplified counterparts in $(SV.ex1)$, $(SV.ex2)$, $(SV.ex3)$, $(SV.Conf.ex4)$, and $(SV.Med.ex4)$. When the true longitudinal model is $(LS.ex4)$, with a time-varying confounder $(W_t)_{t\geq 1}$ affected by the exposure, two possible over-simplified counterparts can be considered, depending on whether $(W_t)_{t\geq 1}$ is mainly considered as a confounder or a mediator.

to "implement" $do(\mathscr{X} = x)$ in the population. In the second counterfactual population, for any profile $\bar{x}^*_{t_0}$ leading to x^*, a proportion $\mathbb{P}(\bar{X}_{t_0} = \bar{x}^*_{t_0} \mid \mathscr{X} = x^*)$ of the individuals undergo the intervention $do(\bar{X}_{t_0} = \bar{x}^*_{t_0})$, which is again a natural way to "implement" $do(\mathscr{X} = x^*)$ in the population. Other averages could be considered, such as weighted averages of longitudinal stratum-specific causal effects. In addition, we shall stress that the interpretability of such quantity is not always straightforward, as was already the case for the weighted averages in Theorem 2 of Section 3.3; we will get back to this point later.

Irrespective of the relevance of the treatment, when only data on $\mathscr{X}$ and $\mathscr{Z}$ are considered or available, practitioners generally (i) overlook the time-varying nature of the exposures, (ii) work under an over-simplified causal model (SV), and (iii) consider the causal effect $ATE_{SV}(x; x^*) = \mathbb{E}_{SV}\left(Y^{\mathscr{X}=x} - Y^{\mathscr{X}=x^*}\right)$, for any $x \neq x^*$, as the causal effect of interest. For example, when the true longitudinal model is model $(LS.ex1)$ given in Figure 3.3, they would implicitly work under model $(SV.ex1)$, while they would typically work under the simplified model $(SV.ex2)$ if the true model is $(LS.ex2)$. Again, there are true longitudinal models under which distinct over-simplified models may be considered in practice. Depending on whether $(W_t)_{t\geq 1}$ is considered to mainly act as a confounder or a mediator under the model $(LS.ex4)$ of Figure 3.3, practitioners would work under either model $(SV.Conf.ex4)$ or model $(SV.Med.ex4)$.

In any case, given an over-simplified model (SV), the causal measure of interest $ATE_{SV}(x; x^*)$ would be estimated in practice and, again, a natural question is whether - and how - this quantity relates to the longitudinal causal effects under the true longitudinal model (LS). Here again, we will use ATE_{SV} when referring to either the causal effect or the quantity estimated in practice in the text. Theorem 3 presents a sufficient condition under which the quantity estimated in practice expresses as a weighted average of stratum specific longitudinal total effects. It is the analogue of Theorem 2 in Section 3.3.1.

Theorem 3. *If condition* $(T3.Cond)$ *below holds*

> *(T3.Cond) There exists some observed* $\mathscr{W} \subset \mathscr{Z}$ *taking its values in* $\Omega_{\mathscr{W}}$*, such that* $(Y^{\mathscr{X}=x} \perp\!\!\!\perp \mathscr{X} \mid \mathscr{W})_{SV}$ *and* $(Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} \perp\!\!\!\perp \bar{X}_{t_0} \mid \mathscr{W})_{LS}$

then the quantity estimated in practice

$$ATE_{SV}(x; x^*) \;\fallingdotseq\; \sum_{w\in\Omega_{\mathscr{W}}} [\mathbb{E}(Y \mid \mathscr{W} = w, \mathscr{X} = x) - \mathbb{E}(Y \mid \mathscr{W} = w, \mathscr{X} = x^*)] \times \mathbb{P}(\mathscr{W} = w),$$

equals

$$\sum_{w\in\Omega_{\mathscr{W}}} \sum_{\substack{\bar{x}_{t_0}\in\{0,1\}^{t_0} \\ \bar{x}^*_{t_0}\in\{0,1\}^{t_0}}} \{ATE_{LS_{|\mathscr{W}=w}}\left(\bar{x}_{t_0}; \bar{x}^*_{t_0}\right) \times \mathbb{P}(\bar{X}_{t_0} = \bar{x}_{t_0} \mid \mathscr{X} = x, \mathscr{W} = w) \times \mathbb{P}(\bar{X}_{t_0} = \bar{x}^*_{t_0} \mid \mathscr{X} = x^*, \mathscr{W} = w) \times \mathbb{P}(\mathscr{W} = w)\}. \tag{3.9}$$

In particular, if condition $(T3.Uncond)$ *below holds*

$$(T3.Uncond) \quad (Y^{\mathscr{X}=x} \perp\!\!\!\perp \mathscr{X})_{SV} \text{ and } (Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} \perp\!\!\!\perp \bar{X}_{t_0})_{LS}$$

then

$$\begin{aligned} ATE_{SV}(\boldsymbol{x};\boldsymbol{x}^*) &\fallingdotseq \mathbb{E}\left(Y \mid \mathscr{X}=\boldsymbol{x}\right) - \mathbb{E}\left(Y \mid \mathscr{X}=\boldsymbol{x}^*\right), \\ &= \sum_{\substack{\bar{x}_{t_0}\in\{0,1\}^{t_0} \\ \bar{x}^*_{t_0}\in\{0,1\}^{t_0}}} \{ATE_{LS}\left(\bar{x}_{t_0};\bar{x}^*_{t_0}\right) \times \mathbb{P}(\bar{X}_{t_0}=\bar{x}_{t_0} \mid \mathscr{X}=\boldsymbol{x}) \\ &\qquad\qquad \times \mathbb{P}(\bar{X}_{t_0}=\bar{x}^*_{t_0} \mid \mathscr{X}=\boldsymbol{x}^*)\}. \end{aligned} \tag{3.10}$$

An analogue of Theorem 1 could be given too: if there exists some observed $\mathscr{W} \subset \mathscr{Z}$ taking values in $\Omega_{\mathscr{W}}$, such that $(Y^{\mathscr{X}=x} \perp\!\!\!\perp \mathscr{X}|\mathscr{W})_{SV}$ and $(Y^{\mathscr{X}=x} \perp\!\!\!\perp \mathscr{X}|\mathscr{W}))_{LS}$, then the quantity estimated in practice equals $ATE_{LS}(\boldsymbol{x},\boldsymbol{x}')$. However, the latter quantity being generally not-well defined from a practical point-of-view unless condition $(Irrel)$ holds, we consider a slightly stronger sufficient condition in Theorem 4 below.

Theorem 4. *Assume that condition* $(Irrel)$ *holds. If, in addition, either condition* $(T3.Cond)$ *or* $(T3.Uncond)$ *holds, then*

$$ATE_{SV}(\boldsymbol{x};\boldsymbol{x}^*) \fallingdotseq ATE_{LS}(\boldsymbol{x};\boldsymbol{x}^*) = ATE_{LS}\left(\bar{x}_{t_0};\bar{x}^*_{t_0}\right),$$

for any $\bar{x}_{t_0}$ *and* $\bar{x}^*_{t_0}$ *leading to to* $\mathscr{X}=\boldsymbol{x}$ *and* $\mathscr{X}=\boldsymbol{x}^*$*, respectively.*

Detailed proofs of Theorems 3 and 4 are given in Appendices 3.B.1 and 3.B.2, respectively. In Section 3.4.2, we illustrate their implications by focusing on a few simple examples.

3.4.2 Examples and illustration of the general results

When the conditions of Theorem 4 are satisfied, the interpretation of the quantity estimated in practice, $ATE_{SV}\left(\boldsymbol{x};\boldsymbol{x}^*\right)$, is straightforward as it equals $ATE_{LS}(\bar{x}_{t_0};\bar{x}^*_{t_0})$ for any $\bar{x}_{t_0}$ and $\bar{x}^*_{t_0}$ leading to $\mathscr{X}=\boldsymbol{x}$ and $\mathscr{X}=\boldsymbol{x}^*$, respectively. However, these conditions are very restrictive. Among the examples presented in Figure 3.3, they are only satisfied under model $(LS.ex1)$, for which the over-simplified counterpart is model $(SV.ex1)$. As made clearer below, condition $(Irrel)$ is not satisfied for models $(LS.ex2)$, $(LS.ex3)$, and $(LS.ex4)$ of Figure 3.3. On the other hand, under model $(LS.ex1)$, versions are irrelevant, and we further have $\left(Y^{\mathscr{X}=x} \perp\!\!\!\perp \mathscr{X} \mid \mathscr{W}\right)_{SV.ex1}$ and $\left(Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} \perp\!\!\!\perp \bar{X}_{t_0} \mid \mathscr{W}\right)_{LS.ex1}$. Therefore, the conditions of Theorem 4 are satisfied, and even if model $(SV.ex1)$ is misspecified ($\mathscr{W}$ is not a confounder for the $(\mathscr{X}-Y)$ relationship under the true model $(LS.ex1)$), the parameter estimated under this over-simplified model retains the parameter of interest $ATE_L(\boldsymbol{x};\boldsymbol{x}^*)$. In other words, observing $\mathscr{X}$ and $\mathscr{W}$ is sufficient to infer the causal effect of $\bar{X}_{t_0}$ on Y under model $(LS.ex1)$.

Below, we will discuss the interpretability of the weighted average in Equations (3.9) and (3.10) when the conditions of Theorem 3 are satisfied. Before that, we shall stress that these conditions are also quite restrictive. They are satisfied in the pure mediation setting, in the absence of further confounding, as depicted in model (*LS.ex2*) of Figure 3.3; see model (*SV.ex2*) for its over-simplified counterpart. First note that, because $\bar{X}_{t_0}$ has an effect on the outcome not only through $\mathscr{X}$ but also through $\mathscr{M}$ under this model, treatment versions are relevant as mentioned above. Nevertheless, $\left(Y^{\mathscr{X}=x} \perp\!\!\!\perp \mathscr{X}\right)_{SV.ex2}$ and $\left(Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} \perp\!\!\!\perp \bar{X}_{t_0}\right)_{LS.ex2}$, so that Theorem 3 ensures that $ATE_{SV}(x; x^*)$ expresses as the weighted average of longitudinal total effects given in Equation (3.10). The conditions of Theorem 3 are still satisfied in the presence of an additional time-invariant pure confounder. But, they are not satisfied anymore if the additional confounder is time-varying, as in model ($LS.ex3$): because of the presence of a time-varying mediator and of a time-varying confounder, $\mathscr{W}$ is no longer sufficient to block all back-door paths between $\bar{X}_{t_0}$ and Y (except if $\bar{W}_{t_0}$ has no direct effect on $\bar{M}_{t_0}$). Consequently, if the true model is ($LS.ex3$), the quantity estimated in practice generally has to be interpreted with caution. See Web Supplementary Material 3.D.1 for more details. Interestingly, this is in sharp contrast with the scenario of model ($LS.ex1$), where only a time-varying pure confounder, and no time-varying pure mediator, was present, and in which case we have already explained that Theorem 4 guaranteed that ATE_{SV} had a clear interpretation. In other words, in the presence of time-varying confounding, the existence of a time-varying mediator is crucial, although it is generally overlooked when the focus is on the estimation of total effects: if there exists a time-varying mediator on top of the time-varying confounder, the conditions of Theorem 3 are not satisfied, and information on summary variables is generally not enough to derive interpretable causal effects.

Another simple example where the conditions of Theorem 3, and *a fortiori*, those of Theorem 4, are not satisfied arises when a time-varying confounder is affected by the exposure of interest, as in Figure 3.3 (*LS.ex4*). First, treatment versions are relevant in this case, since $\bar{X}_{t_0}$ has an effect on the outcome not only through $\mathscr{X}$, but also through $\mathscr{W}$. Moreover, we recall that in this case, two over-simplified models, (*SV.Conf.ex4*) and (*SV.Med.ex4*), may be considered, depending on whether $(W_t)_{t\geq 1}$ is mainly regarded as a confounder or a mediator. Irrespective of the considered over-simplified model, the conditions of Theorem 3 are not satisfied. Indeed, while sequential ignorability holds (more precisely, $(Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} \perp\!\!\!\perp X_1 \mid W_1)_{LS.ex4}$ and $\left(Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} \perp\!\!\!\perp \bar{X}_t \mid \{\bar{W}_t, \bar{X}_{t-1}\}\right)_{LS.ex4}$, for any $t \in [\![2; t_0]\!]$), we do not have $(Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} \perp\!\!\!\perp \bar{X}_{t_0} \mid \mathscr{W})_{LS.ex4}$, and we do not have $(Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} \perp\!\!\!\perp \bar{X}_{t_0})_{LS.ex4}$ either, because $(W_t)_{t>1}$ acts as both a confounder and a mediator in the $(\bar{X}_{t_0} - Y)$ relationship. Therefore, and as detailed in Web Supplementary Material 3.D.2, the quantity estimated under an over-simplified model has to be interpreted with caution if the true longitudinal model is (*L.ex4*), as it generally differs from the causal effects of interest.

We will now provide numerical examples to illustrate the magnitude of these differences. We consider a causal model of the same form as $(LS.ex4)$ in Figure 3.3, with $t_0 = 5$, binary variables X_t and W_t for all $t = 1, \ldots, 5$, and a continuous outcome Y. For any variable U involved in this model, denote the exogeneous variable and structural function corresponding to U by ξ_U and f_U, respectively. We consider the causal model where $\xi_Y \sim \mathcal{N}(0,1)$ while all other exogeneous variables are univariate random variables uniformly distributed on $[0,1]$, and

$$
\begin{aligned}
f_{W_1}(\xi_{W_1}) &= \mathbb{1}\{W_1 \leq 0.1\}, \\
f_{X_1}(W_1, \xi_{X_1}) &= \mathbb{1}\{X_1 \leq \text{expit}(\alpha W_1 + c_{X_1})\}, \\
f_{W_t}\left(\bar{W}_{t-1}, \bar{X}_{t-1}, \xi_{W_t}\right) &= \mathbb{1}\Big\{W_t \leq \text{expit}\big(\gamma \sum_{t'<t} W_{t'} + \rho\alpha X_{t-1} + c_{W_t}\big)\Big\}, \forall t \in [\![2; t_0]\!], \\
f_{X_t}\left(\bar{W}_t, \bar{X}_{t-1}, \xi_{X_t}\right) &= \mathbb{1}\Big\{X_t \leq \text{expit}\big(\alpha \sum_{t'\leq t} W_{t'} + \beta X_{t-1} + c_{X_t}\big)\Big\}, \forall t \in [\![2; t_0]\!], \\
f_Y(\mathscr{X}, \mathscr{W}, \xi_Y) &= \mu_0 + \mu_X \mathscr{X} - \mu_W \mathscr{W} + \xi_Y.
\end{aligned} \tag{3.11}
$$

Here expit$(\cdot)$ denotes the sigmoid function, $\mathbb{1}\{\cdot\}$ denotes the indicator function, $\mathscr{X} = \mathbb{1}\left(\sum_{t=1}^{t_0} X_t \geq 3\right)$ and $\mathscr{W} = \mathbb{1}\left(\sum_{t=1}^{t_0} W_t \geq 3\right)$. Constant terms c_{W_1}, c_{W_t}, and c_{X_t} were chosen so that prevalences of X_t and W_t are about 0.1 for all t and any combination of the parameters α, β, γ and ρ. For instance, we set $c_{W_1} = \text{logit}(0.1) - \dfrac{0.1}{\alpha}$, with $\text{logit}(p) = \log[p/(1-p)]$, for $p \in [0,1]$.

In this model, parameter α governs the strength of the effect of W_t on $X_{t'}$ for $t' \geq t$, while the strength of the effect of X_t on W_{t+1} is governed by the product $\rho\alpha$. The special case $\rho = 0$ corresponds to the scenario where the confounder is not affected by the exposure of interest (pure confounding), while $\alpha = 0$ corresponds to the case where the exposure of interest and the confounder are not causally related (no mediation, no confounding). On the other hand, as parameter ρ increases, we get closer to the pure mediation setting as the effect of the "confounder" on the exposure of interest gets more and more negligible compared to the effect of the exposure on the "confounder". For negative values of parameter α, this simple causal model could be regarded as a simplified version of the causal model describing obesity on the age interval, say, [20-30] (process X_t), physical activity on the same age interval [20-30] (process W_t) and blood pressure at, say, 35 years old (Y).

Under this model, we can derive the analytic expression of (i) $ATE_{LS.ex4}(\bar{x}_{t_0};\, \bar{x}^*_{t_0})$, for any pair of exposure profiles $(\bar{x}_{t_0}; \bar{x}^*_{t_0})$ leading to $\mathscr{X} = 1$ and $\mathscr{X} = 0$, (ii) the weighted average given in Equation (3.8), but also (iii) $ATE_{SV.Conf.ex4}(1;0)$, and (iv) $ATE_{SV.Med.ex4}(1;0)$. Figure 3.4 presents the values of these four quantities for $\alpha \in [-3, 3]$, $\rho \in \{0, 0.1, 0.5, 1, 2, 5, 10\}$ and $\mu_W \in \{0.5, 1, 2\}$. The other parameters were set to $\gamma = \beta = 1$, $\mu_0 = 1$ and $\mu_X = 1$.

In the pure confounding case (when $\rho = 0$), $ATE_{SV.Conf.ex4}(\boldsymbol{x}; \boldsymbol{x}^*)$ equals $ATE_{LS.ex4}(\bar{x}_{t_0};$

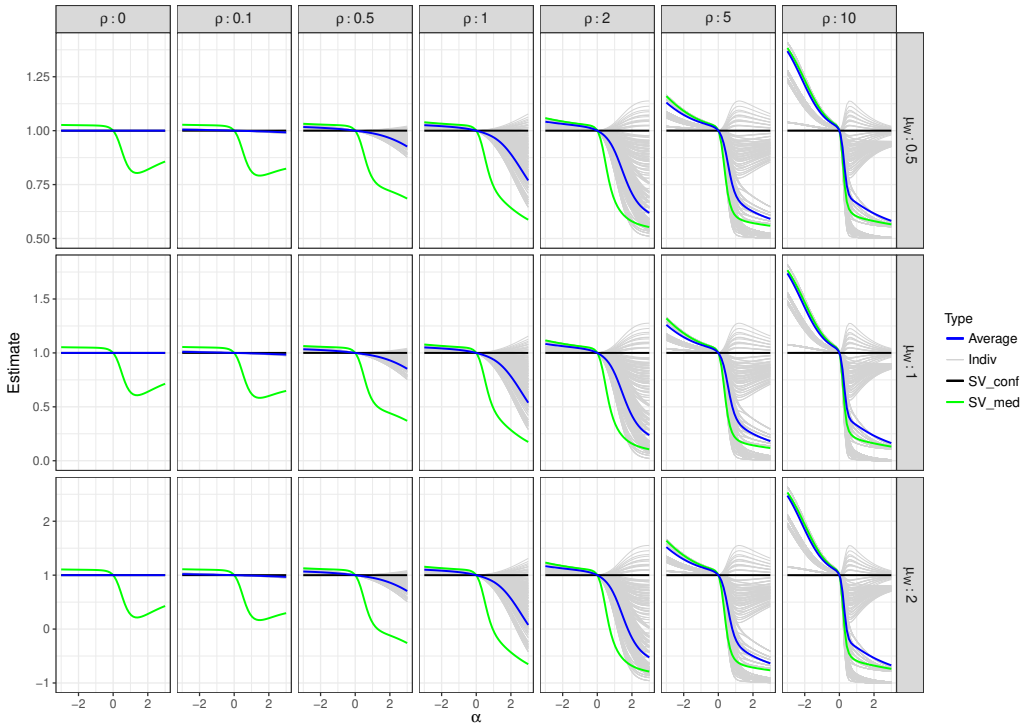


Figure 3.4: Analytic values of $ATE_{SV.Conf}(1;0)$ (in black), $ATE_{SV.Med}(1;0)$ (in green), $ATE_L(\bar{x}_{t_0};\bar{x}^*_{t_0})$ (in grey) for each couple of exposure profiles leading to $\mathscr{X}=1$ and $\mathscr{X}=0$ and the weighted average (3.8) of all these possible comparisons (in blue) under the causal model described in Equation (3.11).

$\bar{x}^*_{t_0}\big)$ for any $\bar{x}_{t_0}$ and $\bar{x}^*_{t_0}$ leading to $\mathscr{X}=x$ and $\mathscr{X}=x^*$, as expected, and thus equals the quantity of interest given in Equation (3.8) as well. This also happens when $\alpha=0$, which corresponds to the "no mediation and no confounding" scenario, in which case $ATE_{SV.Conf}=ATE_{SV.Med}=ATE_L\left(\bar{x}_{t_0};\bar{x}^*_{t_0}\right)$ for any $\bar{x}_{t_0}$ and $\bar{x}^*_{t_0}$ leading to $\mathscr{X}=x$ and $\mathscr{X}=x^*$, and so the weighted average given in Equation (3.8) is also equal to $ATE_{SV.Conf}$. For all other combinations of parameters, both $ATE_{SV.Conf}$ and $ATE_{SV.Med}$ differ from the weighted average of Equation (3.8). When $\rho\in\{0.1,0.5\}$, $(W_t)_{t\geq 1}$ mostly acts as a confounder (and not so much as a mediator), and the difference between $ATE_{SV.Conf}$ and the weighted average is generally limited. As ρ increases, the difference between $ATE_{SV.Conf}$ and the weighted average increases too. Moreover, because the effect of $\mathscr{W}$ on Y is $-\mu_W$, the indirect effect of the exposure process through $(W_t)_t$ is negative for positive α, so that the weighted average can be negative, while $ATE_{SV.Conf}$ suggests a positive association, for some combinations of large values for ρ, α and μ_W. On the other hand, when ρ is large, $(W_t)_{t\geq 1}$ mostly acts as a mediator, and the difference between $ATE_{SV.Med}$ and the weighted average is typically small. It is also noteworthy that the weighted average (3.8) happens to lie between $ATE_{SV.Conf.ex4}(1;0)$ and $ATE_{SV.Med.ex4}(1;0)$ in all the settings presented in Figure 3.4, although it does not hold true in general.

Finally, let us discuss the interpretability of the weighted average of Equation (3.8) (in blue in Figure 3.4), which may or may not equal the quantity estimated in practice

(basically, it equals $ATE_{SV.Conf}$ if $\alpha = 0$ or $\rho = 0$, while it is equal to $ATE_{SV.Med}$ if $\alpha = 0$, and approximately equal to $ATE_{SV.Med}$ if $\alpha \ll \rho$). Figure 3.4 nicely illustrates that the values of the "individual" causal effects $ATE_{LS}\left(\bar{x}_{t_0}; \bar{x}^*_{t_0}\right)$, for different pairs of profiles $\bar{x}_{t_0}$ and $\bar{x}^*_{t_0}$ leading respectively to $\mathscr{X} = 1$ and $\mathscr{X} = 0$, may be quite heterogeneous for some combination of the parameters, while they are more homogeneous for others combinations. For instance, for negative values of α and $\rho \leq 2$, the values of the individual causal effects $ATE_{LS}\left(\bar{x}_{t_0}; \bar{x}^*_{t_0}\right)$ are quite homogeneous. In particular, versions are actually irrelevant for $\rho = 0$ or $\alpha = 0$, and all the individual causal effects $ATE_{LS}\left(\bar{x}_{t_0}; \bar{x}^*_{t_0}\right)$ are equal. In all these cases, the weighted average is then straightforward to interpret. However, the values of the individual causal effects are quite heterogeneous for other combinations of the parameters, especially when both ρ and α are large: in these situations, the weighted average of Equation (3.8), and consequently $ATE_{SV.Med}$, have to be interpreted with caution. This echoes our discussion at the end of Section 3.3.2 where possibly substantially different individual causal effects, such as $ATE_L\big(\mathbf{1}_{t_0}; \mathbf{0}_{t_0}\big)$ and $ATE_L\big((\mathbf{0}_{t_0}, 1); (\mathbf{1}_{t_0-1}, 0)\big)$, could contribute to the weighted average given in Equation (3.6) unless, e.g., some stability assumption held. Actually, a closer inspection into the values of the weights $\mathbb{P}(\bar{X}_{t_0} = \bar{x}_{t_0} \mid \mathscr{X} = \boldsymbol{x}) \times \mathbb{P}(\bar{X}_{t_0} = \bar{x}^*_{t_0} \mid \mathscr{X} = \boldsymbol{x}^*)$ is instructive. For example, consider the case where $\alpha = 3$, $\rho = 10$ and $\mu_W = 2$. In this case, the weighted average in Equation (3.8) is a weighted sum of the three following terms mostly, whose cumulative weight is more than 82%: $ATE_{LS}\big(\mathbf{1}_5; \mathbf{0}_5\big) = -1$, $ATE_{LS}\big((\mathbf{0}_1, \mathbf{1}_4); \mathbf{0}_5\big) = -1$ and $ATE_{LS}\big((\mathbf{0}_2, \mathbf{1}_3); \mathbf{0}_5\big) = 0.8$. Interestingly, although these three causal effects compare the single never-exposed profile to three ever-exposed types of profiles, their values are substantially different. This is again due to the negative indirect effect of the exposure through the $(W_t)_{t\geq 1}$ process. In this particular example, the weighted average of longitudinal causal effects mostly compares ever-exposed profiles to the single never-exposed profile, and can therefore be seen as a causal quantity of interest, even if it is the average of quite different "individual" causal effects. This situation can of course arise as well for the weighted average of Equation (3.7) under the stability assumption in Section 3.3.2.

3.5 Discussion

The longitudinal nature of risk factors is most often overlooked in epidemiology. In this article, we investigated whether causal effects derived when working under simplified, hence generally misspecified, models could still be related to causal effects of potential interest. We focused on two situations regarding exposures: when inference is based on (*i*) their "instantaneous" levels measured at inclusion in the study, and (*ii*) some summary measures of their levels up to inclusion in the study, assuming that these summary measures capture the whole effect of the exposure processes on the outcome. Unsurprisingly, our results are mostly negative, in the sense that the quantity estimated in practice

when working under over-simplified causal models has generally no clear interpretation in terms of longitudinal causal effects of interest, except under very simple longitudinal causal models. Under the conditions of Theorems 1 or Theorem 4, the quantity estimated in practice has a clear interpretation, as it coincides with longitudinal total effects. But, these conditions are very restrictive. Under slightly less restrictive conditions, Theorem 2 and Theorem 3 ensure that the quantity estimated in practice expresses as a weighted average of longitudinal causal effects. But, these conditions are still quite restrictive, and the interpretability of these weighted averages is not always straightforward.

When inference is based on instantaneous levels of exposures measured at inclusion, practitioners should be extremely cautious when interpreting their results as the quantity of interest can generally not be related to any causal effects of interest. A noticeable exception is when a stability assumption holds for the exposure profile and no time-varying confounder is present. In the situation where summary measures are available and capture the whole effect of past levels of exposures, the quantity estimated in practice can be related to causal effects of interest under a few simple causal models. This is the case when the versions of the treatment are irrelevant, and either condition (*T3.Cond*) or (*T3.Uncond*) is verified, as for example in the presence of a time-varying pure confounder only; see model ($LS.ex1$) of Figure 3.3. When the versions are relevant and condition (*T3.Cond*) or (*T3.Uncond*) is verified, the quantity of interest can be expressed as a weighted average of causal effects of interest: this is notably the case in the presence of a time-varying pure mediator only; see model ($LS.ex2$) of Figure 3.3. Moreover, as soon as a time-varying confounder affected by the exposure is present, and/or both time-varying pure mediators and confounders are present, the quantity estimated in practice has to be interpreted with caution since it can generally not be related to any causal effect of interest. We shall stress that even if time-varying pure mediators are generally overlooked when the focus is on total effects, they are likely to exist in most cases. As soon as time-varying confounders exist too, summary variables are no longer sufficient to derive meaningful estimates for total causal effects.

Overall, our results are in line with, and complete, those of previous works, which established the necessity of applying appropriate statistical methods on repeated measurements of exposures when the true causal model is longitudinal (Daniel et al., 2012, Maxwell and Cole, 2007, Maxwell et al., 2011). Even if measurements of exposures are available at baseline only, it would be good practice to still consider the true longitudinal causal model, rather than its over-simplified counterpart. General results on the identifiability of causal effects in the presence of unobserved variables could then be applied (Huang and Valtorta, 2006, Shpitser and Pearl, 2006, Tian and Pearl, 2002) to check whether some longitudinal causal effects of interest can be identified from the available data, even if this will only be the case under very particular and simple causal models. The development of sensitivity analyses may be required for more general models. But,

above all, we believe that forthcoming observational studies should plan the collection of repeated measurements, as a few studies already did, including for biomarkers (Kim et al., 2017). Following these recommendations is likely even more critical when considering time-varying outcomes as in survival analysis, and when targeting causal effects defined on multiplicative scales such as relative risks and odds-ratios.

Disclaimers

Where authors are identified as personnel of the International Agency for Research on Cancer / World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer / World Health Organization.

3.A Appendix: Technical details in the situation where instantaneous levels at inclusion in the study are available

3.A.1 Proof of Theorem 1

Consider a longitudinal model (L) as depicted in Figure 3.2, and assume that there exists $W_{t_0} \subset Z_{t_0}$ taking values in some space $\Omega_{W_{t_0}}$ such that the conditional ignorability condition $Y^{X_{t_0}=x_{t_0}} \perp\!\!\!\perp X_{t_0} \mid W_{t_0}$ holds. Then for any x_{t_0} and $x^*_{t_0}$ in $\{0,1\}$, usual arguments of causal inference (that is, the application of the ignorability condition, and the consistency and positivity conditions) (Pearl, 2000, Robins, 1986, Rosenbaum and Rubin, 1983), yields

$$\begin{aligned}
ATE_L\left(x_{t_0}; x^*_{t_0}\right) &:= \mathbb{E}_L\left(Y^{X_{t_0}=x_{t_0}} - Y^{X_{t_0}=x^*_{t_0}}\right), \\
&= \sum_{w_{t_0} \in \Omega_{W_{t_0}}} \mathbb{E}_L\left(Y^{X_{t_0}=x_{t_0}} - Y^{X_{t_0}=x^*_{t_0}} \mid W_{t_0} = w_{t_0}\right) \times \mathbb{P}(W_{t_0} = w_{t_0}), \\
&= \sum_{w_{t_0} \in \Omega_{W_{t_0}}} \Big[\mathbb{E}_L\Big(Y^{X_{t_0}=x_{t_0}} \mid W_{t_0} = w_{t_0}, X_{t_0} = x_{t_0}\Big) \\
&\quad - \mathbb{E}_L\Big(Y^{X_{t_0}=x^*_{t_0}} \mid W_{t_0} = w_{t_0}, X_{t_0} = x^*_{t_0}\Big)\Big] \times \mathbb{P}(W_{t_0} = w_{t_0}), \\
&= \sum_{w_{t_0} \in \Omega_{W_{t_0}}} \Big[\mathbb{E}\Big(Y \mid W_{t_0} = w_{t_0}, X_{t_0} = x_{t_0}\Big) - \mathbb{E}\Big(Y \mid W_{t_0} = w_{t_0}, X_{t_0} = x^*_{t_0}\Big)\Big] \\
&\quad \times \mathbb{P}(W_{t_0} = w_{t_0}).
\end{aligned}$$

Now, consider an over-simplified model (CS) under which $Y^{X_{t_0}=x_{t_0}} \perp\!\!\!\perp X_{t_0} \mid W_{t_0}$. Then the quantity estimated in practice when working under this over-simplified model is, for

any x_{t_0} and $x^*_{t_0}$ in $\{0,1\}$,

$$\begin{aligned}
ATE_{CS}\left(x_{t_0};x^*_{t_0}\right) &:= \mathbb{E}_{CS}\left(Y^{X_{t_0}=x_{t_0}} - Y^{X_{t_0}=x^*_{t_0}}\right),\\
&= \sum_{w_{t_0}\in\Omega_{W_{t_0}}} \left[\mathbb{E}_{CS}\left(Y^{X_{t_0}=x_{t_0}} \mid W_{t_0}=w_{t_0}\right) - \mathbb{E}_{CS}\left(Y^{X_{t_0}=x^*_{t_0}} \mid W_{t_0}=w_{t_0}\right)\right]\\
&\quad \times \mathbb{P}(W_{t_0}=w_{t_0}),\\
&= \sum_{w_{t_0}\in\Omega_{W_{t_0}}} \Big[\mathbb{E}_{CS}\left(Y^{X_{t_0}=x_{t_0}} \mid W_{t_0}=w_{t_0}, X_{t_0}=x_{t_0}\right)\\
&\quad -\mathbb{E}_{CS}\left(Y^{X_{t_0}=x^*_{t_0}} \mid W_{t_0}=w_{t_0}, X_{t_0}=x^*_{t_0}\right)\Big] \times \mathbb{P}(W_{t_0}=w_{t_0}),\\
&\overset{\circ}{=} \sum_{w_{t_0}\in\Omega_{W_{t_0}}} \left[\mathbb{E}\left(Y \mid W_{t_0}=w_{t_0}, X_{t_0}=x_{t_0}\right) - \mathbb{E}\left(Y \mid W_{t_0}=w_{t_0}, X_{t_0}=x^*_{t_0}\right)\right]\\
&\quad \times \mathbb{P}(W_{t_0}=w_{t_0})\\
&= ATE_L\left(x_{t_0};x^*_{t_0}\right).
\end{aligned}$$

Using similar arguments, if $(Y^{X_{t_0}=x_{t_0}} \perp\!\!\!\perp X_{t_0})_L$ and $(Y^{X_{t_0}=x_{t_0}} \perp\!\!\!\perp X_{t_0})_{CS}$, it follows that

$$\begin{aligned}
ATE_{CS}\left(x_{t_0};x^*_{t_0}\right) &\overset{\circ}{=} \mathbb{E}\Big(Y \mid X_{t_0}=x_{t_0}\Big) - \mathbb{E}\Big(Y \mid X_{t_0}=x^*_{t_0}\Big)\\
&= ATE_L\left(x_{t_0};x^*_{t_0}\right),
\end{aligned}$$

which completes the proof of Theorem 1.

3.A.2 Proof of Theorem 2

Consider again a longitudinal model (L) as depicted in Figure 3.2, and assume that there exists $W_{t_0} \subset Z_{t_0}$ taking values in some space $\Omega_{W_{t_0}}$ such that $Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} \perp\!\!\!\perp \bar{X}_{t_0} \mid W_{t_0}$. Then for any $\bar{x}_{t_0}$ and $\bar{x}^*_{t_0}$ in $\{0,1\}^{t_0}$, usual arguments of causal inference (Pearl, 2000, Robins, 1986, Rosenbaum and Rubin, 1983) yield

$$\begin{aligned}
ATE_L\left(\bar{x}_{t_0};\bar{x}^*_{t_0}\right) &:= \mathbb{E}_L\left(Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} - Y^{\bar{X}_{t_0}=\bar{x}^*_{t_0}}\right),\\
&= \sum_{w_{t_0}\in\Omega_{W_{t_0}}} \left[\mathbb{E}\Big(Y \mid W_{t_0}=w_{t_0}, \bar{X}_{t_0}=\bar{x}_{t_0}\Big) - \mathbb{E}\Big(Y \mid W_{t_0}=w_{t_0}, \bar{X}_{t_0}=\bar{x}^*_{t_0}\Big)\right]\\
&\quad \times \mathbb{P}(W_{t_0}=w_{t_0}).
\end{aligned}$$

Now, consider an over-simplified model (CS) under which $Y^{X_{t_0}=x_{t_0}} \perp\!\!\!\perp X_{t_0} \mid W_{t_0}$. Then the quantity estimated in practice when working under this over-simplified model is, for any x_{t_0} and $x^*_{t_0}$ in $\{0,1\}$,

$$\begin{aligned}
ATE_{CS}\left(x_{t_0};x^*_{t_0}\right) &:= \mathbb{E}_{CS}\left(Y^{X_{t_0}=x_{t_0}} - Y^{X_{t_0}=x^*_{t_0}}\right),\\
&\overset{\circ}{=} \sum_{w_{t_0}\in\Omega_{W_{t_0}}} \left[\mathbb{E}\left(Y \mid W_{t_0}=w_{t_0}, X_{t_0}=x_{t_0}\right) - \mathbb{E}\left(Y \mid W_{t_0}=w_{t_0}, X_{t_0}=x^*_{t_0}\right)\right]\\
&\quad \times \mathbb{P}(W_{t_0}=w_{t_0}).
\end{aligned}$$

But, under model (L), we have, for any x_{t_0} in $\{0,1\}$ and $w_{t_0} \in \Omega_{W_{t_0}}$,

$$\begin{aligned}\mathbb{E}\left(Y \mid W_{t_0}=w_{t_0}, X_{t_0}=x_{t_0}\right) &= \sum_{\bar{x}_{t_0-1}} \mathbb{E}\Big(Y \mid W_{t_0}=w_{t_0}, \bar{X}_{t_0}=\bar{x}_{t_0}\Big) \\ &\quad \times \mathbb{P}(\bar{X}_{t_0-1}=\bar{x}_{t_0-1} \mid X_{t_0}=x_{t_0}, W_{t_0}=w_{t_0}), \\ &= \sum_{\bar{x}_{t_0-1}} \mathbb{E}_L\Big(Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} \mid W_{t_0}=w_{t_0},\Big) \\ &\quad \times \mathbb{P}(\bar{X}_{t_0-1}=\bar{x}_{t_0-1} \mid X_{t_0}=x_{t_0}, W_{t_0}=w_{t_0}),\end{aligned}$$

where the sum is over all possible values of $\bar{X}_{t_0-1}$ in $\{0,1\}^{t_0-1}$. Therefore,

$$\begin{aligned}ATE_{CS}\left(x_{t_0}; x^*_{t_0}\right) \;\fallingdotseq\; \sum_{w_{t_0}\in\Omega_{W_{t_0}}} \sum_{\substack{\bar{x}_{t_0-1} \\ \bar{x}^*_{t_0-1}}} \{&ATE_{L_{|W_{t_0}=w_{t_0}}}\left(\bar{x}_{t_0}; \bar{x}^*_{t_0}\right) \\ &\times \mathbb{P}(\bar{X}_{t_0-1}=\bar{x}_{t_0-1} \mid X_{t_0}=x_{t_0}, W_{t_0}=w_{t_0}) \\ &\times \mathbb{P}(\bar{X}_{t_0-1}=\bar{x}^*_{t_0-1} \mid X_{t_0}=x^*_{t_0}, W_{t_0}=w_{t_0}) \\ &\times \mathbb{P}(W_{t_0}=w_{t_0})\},\end{aligned}$$

which establishes the result under condition $(T2.Cond)$.

The proof of the result under condition $(T2.Uncond)$ follows from similar, but simpler, arguments and is therefore omitted.

3.B Appendix: Technical details in the situation where summary measures of past exposures are available

3.B.1 Proof of Theorem 3

Consider a longitudinal model (LS) as depicted in Figure 3.3, and assume that there exists $\mathscr{W} \subset \mathscr{Z}$ taking its values in some space $\Omega_{\mathscr{W}}$ such that $Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} \perp\!\!\!\perp X_{t_0} \mid \mathscr{W}$. Then, for any $\bar{x}_{t_0}$ and $\bar{x}^*_{t_0}$ in $\{0,1\}^{t_0}$, usual arguments of causal inference (Pearl, 2000, Robins, 1986, Rosenbaum and Rubin, 1983), yield

$$\begin{aligned}ATE_{LS}\left(\bar{x}_{t_0}; \bar{x}^*_{t_0}\right) &= \sum_{w\in\Omega_{\mathscr{W}}} ATE_{LS_{|\mathscr{W}=w}}\left(\bar{x}_{t_0}; \bar{x}^*_{t_0}\right) \times \mathbb{P}(\mathscr{W}=w), \\ &= \sum_{w\in\Omega_{\mathscr{W}}} \mathbb{E}_{LS}\left(Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} - Y^{\bar{X}_{t_0}=\bar{x}^*_{t_0}} \mid \mathscr{W}=w\right) \times \mathbb{P}(\mathscr{W}=w), \\ &= \sum_{w\in\Omega_{\mathscr{W}}} \Big[\mathbb{E}_{LS}\Big(Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} \mid \mathscr{W}=w, \bar{X}_{t_0}=\bar{x}_{t_0}\Big) \\ &\qquad - \mathbb{E}_{LS}\Big(Y^{\bar{X}_{t_0}=\bar{x}^*_{t_0}} \mid \mathscr{W}=w, \bar{X}_{t_0}=\bar{x}^*_{t_0}\Big)\Big] \times \mathbb{P}(\mathscr{W}=w), \\ &= \sum_{w\in\Omega_{\mathscr{W}}} \Big[\mathbb{E}\Big(Y \mid \mathscr{W}=w, \bar{X}_{t_0}=\bar{x}_{t_0}\Big) - \mathbb{E}\Big(Y \mid \mathscr{W}=w, \bar{X}_{t_0}=\bar{x}^*_{t_0}\Big)\Big] \\ &\qquad \times \mathbb{P}(\mathscr{W}=w).\end{aligned}$$

Now, consider an over-simplified model (SV) under which $Y^{\mathscr{X}=x} \perp\!\!\!\perp \mathscr{X} \mid \mathscr{W}$. Then, the quantity estimated in practice when working under this over-simplified model is, for any given $\boldsymbol{x}, \boldsymbol{x}^*$,

$$\begin{aligned} ATE_{SV}\left(\boldsymbol{x};\boldsymbol{x}^*\right) &:= \mathbb{E}_{SV}\left(Y^{\mathscr{X}=x} - Y^{\mathscr{X}=x^*}\right), \\ &\circeq \sum_{w\in\Omega_{\mathscr{W}}} \left[\mathbb{E}\left(Y \mid \mathscr{W}=w, \mathscr{X}=x\right) - \mathbb{E}\left(Y \mid \mathscr{W}=w, \mathscr{X}=x^*\right)\right] \\ &\quad \times \mathbb{P}(\mathscr{W}=w). \end{aligned}$$

But, because $\bar{X}_{t_0}$ d-separates $\mathscr{X}$ and $\mathscr{W}$ under model (LS) (Pearl, 2000, Verma and Pearl, 1988), we have, for any $\bar{x}_{t_0}$ in $\{0,1\}^{t_0}$ and any w in $\Omega_{\mathscr{W}}$,

$$\begin{aligned} \mathbb{E}\Big(Y \mid \mathscr{W}=w, \bar{X}_{t_0}=\bar{x}_{t_0}\Big) &= \sum_x \mathbb{E}\Big(Y \mid \mathscr{W}=w, \mathscr{X}=x, \bar{X}_{t_0}=\bar{x}_{t_0}\Big) \\ &\quad \times \mathbb{P}(\mathscr{X}=x \mid \mathscr{W}=w, \bar{X}_{t_0}=\bar{x}_{t_0}), \\ &= \sum_x \mathbb{E}\Big(Y \mid \mathscr{W}=w, \mathscr{X}=x, \bar{X}_{t_0}=\bar{x}_{t_0}\Big) \\ &\quad \times \mathbb{P}(\mathscr{X}=x \mid \bar{X}_{t_0}=\bar{x}_{t_0}), \\ &= \mathbb{E}\Big(Y \mid \mathscr{W}=w, \mathscr{X}=x, \bar{X}_{t_0}=\bar{x}_{t_0}\Big), \end{aligned}$$

with x corresponding to the value taken by $\mathscr{X}$ when $\bar{X}_{t_0}=\bar{x}_{t_0}$. In other respect, for any $\boldsymbol{x}$, we have

$$\begin{aligned} \mathbb{E}\left(Y \mid \mathscr{W}=w, \mathscr{X}=x\right) &= \sum_{\bar{x}_{t_0}} \mathbb{E}\left(Y \mid \mathscr{W}=w, \mathscr{X}=x, \bar{X}_{t_0}=\bar{x}_{t_0}\right) \\ &\quad \times \mathbb{P}(\bar{X}_{t_0}=\bar{x}_{t_0} \mid \mathscr{W}=w, \mathscr{X}=x), \\ &= \sum_{\bar{x}_{t_0}} \mathbb{E}\left(Y \mid \mathscr{W}=w, \bar{X}_{t_0}=\bar{x}_{t_0}\right) \\ &\quad \times \mathbb{P}(\bar{X}_{t_0}=\bar{x}_{t_0} \mid \mathscr{W}=w, \mathscr{X}=x), \\ &= \sum_{\bar{x}_{t_0}} \mathbb{E}_{LS}\Big(Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} \mid \mathscr{W}=w\Big) \\ &\quad \times \mathbb{P}(\bar{X}_{t_0}=\bar{x}_{t_0} \mid \mathscr{W}=w, \mathscr{X}=x), \end{aligned}$$

where the second equality comes from the fact that $\bar{X}_{t_0}=\bar{x}_{t_0} \Rightarrow \mathscr{X}=x$ for any $\bar{x}_{t_0}$ such that $\mathbb{P}(\bar{X}_{t_0}=\bar{x}_{t_0} \mid \mathscr{W}=w, \mathscr{X}=x) \neq 0$. This finally yields

$$\begin{aligned} ATE_{SV}(x;x^*) \circeq \sum_{w\in\Omega_{\mathscr{W}}} \sum_{\substack{\bar{x}_{t_0} \\ \bar{x}^*_{t_0}}} \{ & ATE_{LS_{|\mathscr{W}=w}}\left(\bar{x}_{t_0};\bar{x}^*_{t_0}\right) \times \mathbb{P}(\bar{X}_{t_0}=\bar{x}_{t_0} \mid \mathscr{X}=x, \mathscr{W}=w) \\ & \times \mathbb{P}(\bar{X}_{t_0}=\bar{x}^*_{t_0} \mid \mathscr{X}=x^*, \mathscr{W}=w) \\ & \times \mathbb{P}(\mathscr{W}=w)\}, \end{aligned}$$

where the sums are over all $\bar{x}_{t_0}$ and $\bar{x}^*_{t_0}$ in $\{0,1\}^{t_0}$ such that $\mathbb{P}(\bar{X}_{t_0}=\bar{x}_{t_0} \mid \mathscr{W}=w, \mathscr{X}=x)$ and $\mathbb{P}(\bar{X}_{t_0}=\bar{x}^*_{t_0} \mid \mathscr{W}=w, \mathscr{X}=x^*)$, respectively, are not null.

The proof of the result under condition $(T3.Uncond)$ follows from similar, but simpler, arguments and is therefore omitted.

3.B.2 Proof of Theorem 4

First consider a model (LS) as depicted in Figure 3.3, and assume that the versions of the treatment are irrelevant, and that there exists $\mathscr{W} \subset \mathscr{Z}$ such that $\left(Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} \perp\!\!\!\perp \bar{X}_{t_0} \mid \mathscr{W}\right)_{LS}$ and $\left(Y^{\mathscr{X}=x} \perp\!\!\!\perp \mathscr{X} \mid \mathscr{W}\right)_{SV}$. Consider any given $x \neq x^*$; for any $\bar{x}_{t_0}$ such that $\bar{X}_{t_0} = \bar{x}_{t_0} \Rightarrow \mathscr{X} = x$, we have $Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} = Y^{\mathscr{X}=x}$. Therefore, for such $\bar{x}_{t_0}$, we *a fortiori* have $\mathbb{E}\left(Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} \mid \mathscr{W} = w\right) = \mathbb{E}\left(Y^{\mathscr{X}=x} \mid \mathscr{W} = w\right)$ for any w in Ω_w. As a result, for any w in Ω_w and any $\bar{x}_{t_0}$ and $\bar{x}^*_{t_0}$ leading respectively to $\mathscr{X} = x$ and $\mathscr{X} = x^*$, $ATE_{LS_{|\mathscr{W}=w}}(\bar{x}_{t_0}; \bar{x}^*_{t_0}) = ATE_{LS_{|\mathscr{W}=w}}(x; x^*)$. According to the result of Theorem 3, we finally have

$$\begin{aligned}
ATE_{SV}(x; x^*) \;\fallingdotseq\; & \sum_{w \in \Omega_{\mathscr{W}}} \sum_{\substack{\bar{x}_{t_0} \\ \bar{x}^*_{t_0}}} \{ATE_{LS_{|\mathscr{W}=w}}(x; x^*) \times \mathbb{P}(\bar{X}_{t_0} = \bar{x}_{t_0} \mid \mathscr{X} = x, \mathscr{W} = w) \\
& \qquad \times \mathbb{P}(\bar{X}_{t_0} = \bar{x}^*_{t_0} \mid \mathscr{X} = x^*, \mathscr{W} = w) \\
& \qquad \times \mathbb{P}(\mathscr{W} = w)\}, \\
= & \sum_{w \in \Omega_{\mathscr{W}}} ATE_{LS_{|\mathscr{W}=w}}(x; x^*) \times \mathbb{P}(\mathscr{W} = w), \\
= \;& ATE_{LS}(x; x^*) \\
= \;& ATE_{LS}(\bar{x}_{t_0}; \bar{x}^*_{t_0}).
\end{aligned}$$

In the last equality, $\bar{x}_{t_0}$ and $\bar{x}^*_{t_0}$ are two profiles leading to $\mathscr{X} = x$ and $\mathscr{X} = x^*$, respectively.

The proof of the result under condition $(T3.Uncond)$ follows from similar, but simpler, arguments and is therefore omitted.

(L) (CS)

Figure 3.5: (L) Longitudinal model with time-varying exposure and time-varying confounder not affected by the exposure. (CS) Over-simplified cross-sectional model associated with the longitudinal model given in Figure 3.5 (L).

3.C Web Supplementary Material: Extensions for the situation where instantaneous levels at inclusion in the study are available

3.C.1 In the presence of time-varying pure confounder

Consider the configuration of Figure 3.2 $(L.ex2)$ in the Main Document, which is recalled in Figure 3.5 (L) for convenience. It corresponds to the case where the model involves a time-varying pure confounder $(W_t)_t$. The model of Figure 3.5 (CS) is the corresponding over-simplified version. We denote the space in which $\bar{W}_{t_0}$ takes its values by $\Omega_{\bar{W}_{t_0}}$. We have $(Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} \perp\!\!\!\perp \bar{X}_{t_0} \mid \bar{W}_{t_0})_L$, but, in general $(Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} \not\perp\!\!\!\perp \bar{X}_{t_0} \mid W_{t_0})_L$: so, conditions of Theorem 2 given in Section 3.3.1 in the Main Document are not satisfied.

We can still have a closer inspection on the quantity that would be estimated in practice, when working under the over-simplified model (CS). Because $(Y^{X_{t_0}=x} \perp\!\!\!\perp X_{t_0} \mid W_{t_0})_{CS}$, it follows from standard arguments (Pearl, 2000, Robins, 1986, Rosenbaum and Rubin, 1983) that

$$\begin{aligned} ATE_{CS} &= \mathbb{E}_{CS}\left(Y^{X_{t_0}=1} - Y^{X_{t_0}=0}\right) \\ &\circeq \sum_{\bar{w}_{t_0} \in \Omega_{\bar{W}_{t_0}}} \left\{\mathbb{E}\left(Y \mid X_{t_0}=1, W_{t_0}=w_{t_0}\right) - \mathbb{E}\left(Y \mid X_{t_0}=0, W_{t_0}=w_{t_0}\right)\right\} \times \mathbb{P}(W_{t_0}=w_{t_0}). \end{aligned}$$

Moreover, $(Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} \perp\!\!\!\perp \bar{X}_{t_0} \mid \bar{W}_{t_0})_L$, so that, for any x_{t_0} in $\{0,1\}$ and w_{t_0} in $\Omega_{W_{t_0}}$, we have

$$\begin{aligned} &\mathbb{E}(Y \mid X_{t_0}=x_{t_0}, W_{t_0}=w_{t_0}) \\ &= \sum_{\bar{w}_{t_0-1} \in \Omega_{\bar{W}_{t_0-1}}} \sum_{\bar{x}_{t_0-1}} \mathbb{E}\left(Y \mid \bar{X}_{t_0-1}=\bar{x}_{t_0-1}, X_{t_0}=x_{t_0}, \bar{W}_{t_0-1}=\bar{w}_{t_0-1}, W_{t_0}=w_{t_0}\right) \\ &\qquad \times \mathbb{P}\left(\bar{W}_{t_0-1}=\bar{w}_{t_0-1} \mid X_{t_0}=x_{t_0}, W_{t_0}=w_{t_0}\right) \\ &\qquad \times \mathbb{P}\left(\bar{X}_{t_0-1}=\bar{x}_{t_0-1} \mid X_{t_0}=x_{t_0}, \bar{W}_{t_0-1}=\bar{w}_{t_0-1}, W_{t_0}=w_{t_0}\right), \end{aligned}$$

$$= \sum_{\bar{w}_{t_0-1} \in \Omega_{\bar{W}_{t_0-1}}} \sum_{\bar{x}_{t_0-1}} \mathbb{E}_L \left(Y^{\bar{X}_{t_0-1}=\bar{x}_{t_0-1}, X_{t_0}=x_{t_0}} \mid \bar{W}_{t_0-1} = \bar{w}_{t_0-1}, W_{t_0} = w_{t_0} \right)$$
$$\times \mathbb{P} \left(\bar{W}_{t_0-1} = \bar{w}_{t_0-1} \mid X_{t_0} = x_{t_0}, W_{t_0} = w_{t_0} \right)$$
$$\times \mathbb{P} \left(\bar{X}_{t_0-1} = \bar{x}_{t_0-1} \mid X_{t_0} = x_{t_0}, \bar{W}_{t_0-1} = \bar{w}_{t_0-1}, W_{t_0} = w_{t_0} \right),$$

where the sums are over all possible values of $\bar{X}_{t_0-1}$ in $\{0,1\}^{t_0-1}$.

As a result

$$\begin{aligned} ATE_{CS} \;\; \fallingdotseq \sum_{\bar{w}_{t_0} \in \Omega_{\bar{W}_{t_0}}} \sum_{\bar{x}_{t_0-1}} \sum_{\bar{x}^*_{t_0-1}} \Bigg[& \mathbb{E}_L \left(Y^{\bar{X}_{t_0-1}=\bar{x}_{t_0-1}, X_{t_0}=1} \mid \bar{W}_{t_0} = \bar{w}_{t_0} \right) \\ & \times \mathbb{P} \left(\bar{W}_{t_0-1} = \bar{w}_{t_0-1} \mid X_{t_0} = 1, W_{t_0} = w_{t_0} \right) \\ & - \mathbb{E}_L \left(Y^{\bar{X}_{t_0-1}=\bar{x}^*_{t_0-1}, X_{t_0}=0} \mid \bar{W}_{t_0} = \bar{w}_{t_0} \right) \\ & \times \mathbb{P} \left(\bar{W}_{t_0-1} = \bar{w}_{t_0-1} \mid X_{t_0} = 0, W_{t_0} = w_{t_0} \right) \Bigg] \\ & \times \mathbb{P}(W_{t_0} = w_{t_0}) \\ & \times \mathbb{P} \left(\bar{X}_{t_0-1} = \bar{x}_{t_0-1} \mid X_{t_0} = 1, \bar{W}_{t_0} = \bar{w}_{t_0} \right) \\ & \times \mathbb{P} \left(\bar{X}_{t_0-1} = \bar{x}^*_{t_0-1} \mid X_{t_0} = 0, \bar{W}_{t_0} = \bar{w}_{t_0} \right). \end{aligned}$$

Then because the terms $\mathbb{P} \left(\bar{W}_{t_0-1} = \bar{w}_{t_0-1} \mid X_{t_0} = 1, W_{t_0} = w_{t_0} \right)$ and $\mathbb{P}\big(\bar{W}_{t_0-1} = \bar{w}_{t_0-1} \mid X_{t_0} = 0, W_{t_0} = w_{t_0} \big)$ are generally different, ATE_{CS} cannot be expressed in terms of any sensible longitudinal (or stratum-specific longitudinal) total effect measures. Therefore, ATE_{CS} has to be interpreted with caution under this causal model, as its meaning remains unclear.

Moreover, the stability assumption for the exposure (as defined in Section 3.3.2 in the Main Document) does not help here. Indeed, under this assumption, we have

$$\begin{aligned} ATE_{CS} \;\; \fallingdotseq \sum_{\bar{w}_{t_0} \in \Omega_{\bar{W}_{t_0}}} \sum_{\bar{x}_{t_0-1}} \sum_{i=0}^{t_0-1} \Bigg[& \mathbb{E}_L \left(Y^{\bar{X}_{t_0-1}=(\mathbf{0}_i, \mathbf{1}_{t_0-i})} \mid \bar{W}_{t_0-1} = \bar{w}_{t_0-1}, W_{t_0} = w_{t_0} \right) \\ & \times \mathbb{P} \left(\bar{W}_{t_0-1} = \bar{w}_{t_0-1} \mid X_{t_0} = 1, W_{t_0} = w_{t_0} \right) \\ & - \mathbb{E}_L \left(Y^{\bar{X}_{t_0-1}=\mathbf{0}_{t_0}} \mid \bar{W}_{t_0-1} = \bar{w}_{t_0-1}, W_{t_0} = w_{t_0} \right) \\ & \times \mathbb{P} \left(\bar{W}_{t_0-1} = \bar{w}_{t_0-1} \mid X_{t_0} = 0, W_{t_0} = w_{t_0} \right) \Bigg] \times \mathbb{P}(W_{t_0} = w_{t_0}) \\ & \times \mathbb{P} \left(\bar{X}_{t_0-1} = (\mathbf{0}_i, \mathbf{1}_{t_0-i-1}) \mid X_{t_0} = 1, \bar{W}_{t_0-1} = \bar{w}_{t_0-1}, W_{t_0} = w_{t_0} \right), \end{aligned}$$

which, again, cannot be expressed it in terms of longitudinal total effects or longitudinal stratum-specific longitudinal total effects. One last remark is that assuming that the stability assumption holds for both the exposure and the time-varying confounder does not help either. For example, under this "double" stability assumption, we have, in the

(LS) (SV)

Figure 3.6: (LS) Longitudinal model with a time-varying exposure, a time-varying pure confounder, and a time-varying pure mediator, which all affect the outcome through summary variables only. (SV) Corresponding over-simplified model.

particular case of a univariate binary confounder,

$$
\begin{aligned}
ATE_{CS} \Bumpeq \sum_{i=0}^{t_0-1} \Bigg\{ \sum_{j=0}^{t_0-1} \Bigg[& \mathbb{E}_L\left(Y^{\bar{X}_{t_0}=(\mathbf{0}_i,\mathbf{1}_{t_0-i})} \mid \bar{W}_{t_0} = (\mathbf{0}_j, \mathbf{1}_{t_0-j})\right) \\
& \times \mathbb{P}\left(\bar{W}_{t_0-1} = (\mathbf{0}_j, \mathbf{1}_{t_0-j-1}) \mid X_{t_0} = 1, W_{t_0} = 1\right) \\
& - \mathbb{E}_L\left(Y^{\bar{X}_{t_0}=\mathbf{0}_{t_0}} \mid \bar{W}_{t_0} = (\mathbf{0}_j, \mathbf{1}_{t_0-j})\right) \\
& \times \mathbb{P}\left(\bar{W}_{t_0-1} = (\mathbf{0}_j, \mathbf{1}_{t_0-j-1}) \mid X_{t_0} = 0, W_{t_0} = 1\right) \Bigg] \times \mathbb{P}(W_{t_0} = 1) \\
& \times \mathbb{P}\left(\bar{X}_{t_0-1} = (\mathbf{0}_i, \mathbf{1}_{t_0-i-1}) \mid X_{t_0} = 1, \bar{W}_{t_0} = (\mathbf{0}_j, \mathbf{1}_{t_0-j})\right) \\
& + \left[\mathbb{E}_L\left(Y^{\bar{X}_{t_0}=(\mathbf{0}_i,\mathbf{1}_{t_0-i})} \mid \bar{W}_{t_0} = \mathbf{0}_{t_0}\right) - \mathbb{E}_L\left(Y^{\bar{X}_{t_0}=\mathbf{0}_{t_0}} \mid \bar{W}_{t_0} = \mathbf{0}_{t_0}\right)\right] \\
& \times \mathbb{P}(W_{t_0} = 0) \times \mathbb{P}\left(\bar{X}_{t_0-1} = (\mathbf{0}_i, \mathbf{1}_{t_0-i-1}) \mid X_{t_0} = 1, \bar{W}_{t_0} = \mathbf{0}_{t_0}\right) \Bigg\},
\end{aligned}
$$

which still cannot be expressed it in terms of longitudinal total effects or longitudinal stratum-specific longitudinal total effects.

3.D Web Supplementary Material: Extensions for the situation where summary measures of past exposures are available

3.D.1 In the presence of time-varying pure mediator and time-varying pure confounder

We now turn our attention to the setting of Figure 3.3 (*LS.ex3*) in the Main Document, which is recalled in Figure 3.6 (L) for convenience. It corresponds to the case where the model involves both a time-varying pure confounder $(W_t)_t$ and a time-varying pure mediator $(M_t)_t$. The corresponding over-simplified model is given in Figure 3.6 (SV). We denote the space in which $\bar{W}_{t_0}$ takes its values by $\Omega_{\bar{W}_{t_0}}$.

Because the exposure of interest has an effect on the outcome through $\mathscr{X}$ and $\mathscr{M}$,

this is an example of a compound treatment where versions are relevant: quantities $ATE_{LS}(\bar{x}_{t_0};\bar{x}^*_{t_0})$ for different pairs of exposure profiles $(\bar{x}_{t_0},\bar{x}^*_{t_0})$ leading to $\mathscr{X}=x$ and $\mathscr{X}=x^*$ are typically different. Moreover, because both exposure and confounder processes affect the mediator under this model, we generally have $(Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} \not\perp\!\!\!\perp \bar{X}_{t_0} \mid \mathscr{W})_{LS}$: so, the conditions of Theorem 3 given in Section 3.4.1 in the Main Document are not satisfied.

We can still have a closer inspection on the quantity $ATE_{SV}(x;x^*)=\mathbb{E}_{SV}\big(Y^{\mathscr{X}=x}-Y^{\mathscr{X}=x^*}\big)$, for any given $x\neq x^*$, that would be targeted when working under the over-simplified model Figure 3.6 (SV). Because $(Y \perp\!\!\!\perp \mathscr{X} \mid \mathscr{W})_{SV}$ and $\mathscr{W}$ takes values in $\Omega_{\mathscr{W}}$, it follows from standard arguments (Pearl, 2000, Robins, 1986, Rosenbaum and Rubin, 1983) that

$$ATE_{SV.Conf}(x;x^*) \Bumpeq \sum_{w\in\Omega_{\mathscr{W}}} [\mathbb{E}(Y \mid \mathscr{W}=w,\mathscr{X}=x)-\mathbb{E}(Y \mid \mathscr{W}=w,\mathscr{X}=x^*)] \times \mathbb{P}(\mathscr{W}=w).$$

Moreover, because $(Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} \perp\!\!\!\perp \bar{X}_{t_0} \mid \bar{W}_{t_0})_{LS}$ we have, for any $\bar{x}_{t_0}$ in $\{0,1\}^{t_0}$,

$$\begin{aligned}
\mathbb{E}_{LS}\left(Y^{\bar{X}_{t_0}=\bar{x}_{t_0}}\right) &= \sum_{\bar{w}_{t_0}\in\Omega_{\bar{W}_{t_0}}} \mathbb{E}_{LS}\left(Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} \mid \bar{W}_{t_0}=\bar{w}_{t_0}\right)\times\mathbb{P}(\bar{W}_{t_0}=\bar{w}_{t_0}),\\
&= \sum_{\bar{w}_{t_0}\in\Omega_{\bar{W}_{t_0}}} \mathbb{E}\left(Y \mid \bar{X}_{t_0}=\bar{x}_{t_0},\bar{W}_{t_0}=\bar{w}_{t_0}\right)\times\mathbb{P}(\bar{W}_{t_0}=\bar{w}_{t_0}),\\
&= \sum_{\bar{w}_{t_0}\in\Omega_{\bar{W}_{t_0}}}\sum_{x}\sum_{w\in\Omega_{\mathscr{W}}} \mathbb{E}\left(Y \mid \bar{X}_{t_0}=\bar{x}_{t_0},\bar{W}_{t_0}=\bar{w}_{t_0},\mathscr{X}=x,\mathscr{W}=w\right)\\
&\qquad\times\mathbb{P}(\bar{W}_{t_0}=\bar{w}_{t_0})\times\mathbb{P}(\mathscr{X}=x \mid \bar{X}_{t_0}=\bar{x}_{t_0})\\
&\qquad\times\mathbb{P}(\mathscr{W}=w \mid \bar{W}_{t_0}=\bar{w}_{t_0}),\\
&= \sum_{\bar{w}_{t_0}\in\Omega_{\bar{W}_{t_0}}} \mathbb{E}\left(Y \mid \bar{X}_{t_0}=\bar{x}_{t_0},\bar{W}_{t_0}=\bar{w}_{t_0},\mathscr{X}=x,\mathscr{W}=w\right)\times\mathbb{P}(\bar{W}_{t_0}=\bar{w}_{t_0}),
\end{aligned}$$

with x the value taken by $\mathscr{X}$ when $\bar{X}_{t_0}=\bar{x}_{t_0}$, and w the value taken by $\mathscr{W}$ when $\bar{W}_{t_0}=\bar{w}_{t_0}$. Then, for any x and w, we have

$$\begin{aligned}
\mathbb{E}(Y \mid \mathscr{X}=x,\mathscr{W}=w) &= \sum_{\bar{x}_{t_0}}\sum_{\bar{w}_{t_0}\in\Omega_{\bar{W}_{t_0}}} \mathbb{E}\left(Y \mid \bar{X}_{t_0}=\bar{x}_{t_0},\bar{W}_{t_0}=\bar{w}_{t_0},\mathscr{X}=x,\mathscr{W}=w\right)\\
&\qquad\times\mathbb{P}(\bar{X}_{t_0}=\bar{x}_{t_0},\bar{W}_{t_0}=\bar{w}_{t_0} \mid \mathscr{X}=x,\mathscr{W}=w),\\
&= \sum_{\bar{x}_{t_0}}\sum_{\bar{w}_{t_0}\in\Omega_{\bar{W}_{t_0}}} \mathbb{E}_{LS}\left(Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} \mid \bar{W}_{t_0}=\bar{w}_{t_0}\right)\\
&\qquad\times\mathbb{P}(\bar{X}_{t_0}=\bar{x}_{t_0},\bar{W}_{t_0}=\bar{w}_{t_0} \mid \mathscr{X}=x,\mathscr{W}=w).
\end{aligned}$$

Because of the term $\mathbb{P}(\bar{X}_{t_0}=\bar{x}_{t_0},\bar{W}_{t_0}=\bar{w}_{t_0} \mid \mathscr{X}=x,\mathscr{W}=w)$, the sums in the equation above are restricted over the values $\bar{x}_{t_0}$ and $\bar{w}_{t_0}$ such that $\mathscr{X}=x$ and $\mathscr{W}=w$. Therefore,

we have

$$ATE_{SV}(\boldsymbol{x};\boldsymbol{x}^*) \;\fallingdotseq \sum_{\boldsymbol{w}\in\Omega_{\boldsymbol{\mathscr{W}}}} \sum_{\bar{w}_{t_0}\in\Omega_{\bar{W}_{t_0}}} \Bigg[\sum_{\bar{x}_{t_0}} \mathbb{E}_{LS}\left(Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} \mid \bar{W}_{t_0}=\bar{w}_{t_0}\right) \times \mathbb{P}(\bar{X}_{t_0}=\bar{x}_{t_0} \mid \boldsymbol{\mathscr{X}}=\boldsymbol{x}, \bar{W}_{t_0}=\bar{w}_{t_0}) \times \mathbb{P}(\bar{W}_{t_0}=\bar{w}_{t_0} \mid \boldsymbol{\mathscr{X}}=\boldsymbol{x}, \boldsymbol{\mathscr{W}}=\boldsymbol{w}) - \sum_{\bar{x}^*_{t_0}} \mathbb{E}_{LS}\left(Y^{\bar{X}_{t_0}=\bar{x}^*_{t_0}} \mid \bar{W}_{t_0}=\bar{w}_{t_0}\right) \times \mathbb{P}(\bar{X}_{t_0}=\bar{x}^*_{t_0} \mid \boldsymbol{\mathscr{X}}=\boldsymbol{x}^*, \bar{W}_{t_0}=\bar{w}_{t_0}) \times \mathbb{P}(\bar{W}_{t_0}=\bar{w}_{t_0} \mid \boldsymbol{\mathscr{X}}=\boldsymbol{x}^*, \boldsymbol{\mathscr{W}}=\boldsymbol{w}) \Bigg] \times \mathbb{P}(\boldsymbol{\mathscr{W}}=\boldsymbol{w}).$$

Then because the terms $\mathbb{P}(\bar{W}_{t_0}=\bar{w}_{t_0} \mid \boldsymbol{\mathscr{X}}=\boldsymbol{x}, \boldsymbol{\mathscr{W}}=\boldsymbol{w})$ and $\mathbb{P}(\bar{W}_{t_0}=\bar{w}_{t_0} \mid \boldsymbol{\mathscr{X}}=\boldsymbol{x}^*, \boldsymbol{\mathscr{W}}=\boldsymbol{w})$ are generally different, $ATE_{SV}(\boldsymbol{x};\boldsymbol{x}^*)$ cannot be expressed in terms of any sensible longitudinal (or stratum-specific longitudinal) total effect measures.

We shall further stress that considering the stratum-specific causal effect,

$$ATE_{SV_{|\boldsymbol{\mathscr{W}}=\boldsymbol{w}}}(\boldsymbol{x};\boldsymbol{x}^*) \;=\; \mathbb{E}_{SV}\big(Y^{\boldsymbol{\mathscr{X}}=\boldsymbol{x}} - Y^{\boldsymbol{\mathscr{X}}=\boldsymbol{x}^*} \mid \boldsymbol{\mathscr{W}}=\boldsymbol{w}\big),$$

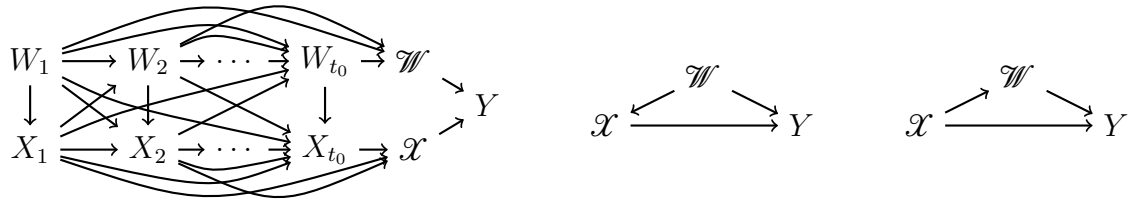
does not help here. Indeed, it can be shown that

$$ATE_{SV_{|\boldsymbol{\mathscr{W}}=\boldsymbol{w}}}(\boldsymbol{x};\boldsymbol{x}^*) \fallingdotseq \sum_{\bar{w}_{t_0}\in\Omega_{\bar{W}_{t_0}}} \Bigg[\sum_{\bar{x}_{t_0}} \mathbb{E}_{LS}\left(Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} \mid \bar{W}_{t_0}=\bar{w}_{t_0}\right) \times \mathbb{P}(\bar{X}_{t_0}=\bar{x}_{t_0} \mid \boldsymbol{\mathscr{X}}=\boldsymbol{x}, \bar{W}_{t_0}=\bar{w}_{t_0}) \times \mathbb{P}(\bar{W}_{t_0}=\bar{w}_{t_0} \mid \boldsymbol{\mathscr{X}}=\boldsymbol{x}, \boldsymbol{\mathscr{W}}=\boldsymbol{w}) - \sum_{\bar{x}^*_{t_0}} \mathbb{E}_{LS}\left(Y^{\bar{X}_{t_0}=\bar{x}^*_{t_0}} \mid \bar{W}_{t_0}=\bar{w}_{t_0}\right) \times \mathbb{P}(\bar{X}_{t_0}=\bar{x}^*_{t_0} \mid \boldsymbol{\mathscr{X}}=\boldsymbol{x}^*, \bar{W}_{t_0}=\bar{w}_{t_0}) \times \mathbb{P}(\bar{W}_{t_0}=\bar{w}_{t_0} \mid \boldsymbol{\mathscr{X}}=\boldsymbol{x}^*, \boldsymbol{\mathscr{W}}=\boldsymbol{w}) \Bigg].$$

Again, because the terms $\mathbb{P}(\bar{W}_{t_0}=\bar{w}_{t_0} \mid \boldsymbol{\mathscr{X}}=\boldsymbol{x}, \boldsymbol{\mathscr{W}}=\boldsymbol{w})$ and $\mathbb{P}(\bar{W}_{t_0}=\bar{w}_{t_0} \mid \boldsymbol{\mathscr{X}}=\boldsymbol{x}^*, \boldsymbol{\mathscr{W}}=\boldsymbol{w})$ are generally different, this quantity cannot be expressed in terms of any sensible longitudinal (or stratum-specific longitudinal) total effect measures.

3.D.2 In the presence of time-varying confounder affected by the exposure

Finally, consider the configuration of Figure 3.3 (*LS.ex4*) in the Main Document, which is recalled in Figure 3.7 (*LS*) for convenience. It corresponds to the case where the model involves a time-varying confounder $(W_t)_t$, which is affected by the exposure. We denote the space in which $\bar{W}_{t_0}$ takes its values by $\Omega_{\bar{W}_{t_0}}$, and by $\Omega_{\boldsymbol{\mathscr{W}}}$ the space in which $\boldsymbol{\mathscr{W}}$ takes



(LS) $(SV.Conf)$ $(SV.Med)$

Figure 3.7: (LS) Longitudinal model with a time-varying exposure and a time-varying confounder affected by the exposure, which both affect the outcome through some summary variables. $(SV.Conf)$ Corresponding over-simplified model if $(W_t)_t$ is mainly considered as a confounder. $(SV.Med)$ Corresponding over-simplified model if $(W_t)_t$ is mainly considered as a mediator.

its values.

Because $(W_t)_{t>1}$ acts as both a confounder and a mediator in the $(\bar{X}_{t_0} - Y)$ relationship, the exposure of interest has an effect on the outcome through $\mathscr{X}$ and $\mathscr{W}$. Therefore, this is another example of a compound treatment where versions are relevant: quantities $ATE_{LS}(\bar{x}_{t_0}; \bar{x}^*_{t_0})$ for different pairs of exposure profiles $(\bar{x}_{t_0}, \bar{x}^*_{t_0})$ leading to $\mathscr{X} = x$ and $\mathscr{X} = x^*$ are typically different. In addition, the ignorability condition $(Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} \perp\!\!\!\perp \bar{X}_{t_0} \mid \mathscr{W})_{LS}$ does not hold, so that the conditions of Theorem 3 given in Section 3.4.1 in the Main Document are not satisfied.

We can still have a closer inspection on the quantity that would be estimated in practice, when working under an over-simplified causal model. If only data on Y, $\mathscr{X}$ and $\mathscr{W}$ is available, two different over-simplified models might be considered in practice: $(SV.Conf)$ or $(SV.Med)$ in Figure 3.7.

First consider the case where $(SV.Conf)$ is (wrongly) considered as the true model. As $(Y^{\mathscr{X}=x} \perp\!\!\!\perp \mathscr{X}|\mathscr{W})_{SV.Conf}$ and, it follows from standard arguments (Pearl, 2000, Robins, 1986, Rosenbaum and Rubin, 1983) that, for any given x, x^*,

$$ATE_{SV.Conf}(x; x^*) \doteqdot \sum_{w \in \Omega_{\mathscr{W}}} [\mathbb{E}(Y \mid \mathscr{W} = w, \mathscr{X} = x) - \mathbb{E}(Y \mid \mathscr{W} = w, \mathscr{X} = x^*)] \times \mathbb{P}(\mathscr{W} = w).$$

But, for any x and w, we have, on the one hand,

$$\mathbb{E}(Y \mid \mathscr{X} = x, \mathscr{W} = w) = \mathbb{E}\left(Y \mid \bar{X}_{t_0} = \bar{x}_{t_0}, \bar{W}_{t_0} = \bar{w}_{t_0}\right),$$

for any $\bar{x}_{t_0}$ and $\bar{w}_{t_0}$ leading to $\mathscr{X} = x$ and $\mathscr{W} = w$, respectively, and, on the other hand

$$\mathbb{P}(\mathscr{W} = w) = \sum_{\bar{w}_{t_0}/\mathscr{W}=w} \mathbb{P}(\bar{W}_{t_0} = \bar{w}_{t_0}),$$

where the sum is over all possible values $\bar{w}_{t_0}$ of $\bar{W}_{t_0}$ in $\Omega_{\bar{W}_{t_0}}$ leading to $\mathscr{W} = w$. As a

result

$$
\begin{aligned}
ATE_{SV.Conf}\left(\boldsymbol{x};\boldsymbol{x}^*\right) \fallingdotseq & \sum_{w\in\Omega_{\mathscr{W}}} \sum_{\substack{\bar{w}_{t_0}\in\Omega_{\bar{W}_{t_0}} \\ /\mathscr{W}=w}} \Big[\mathbb{E}\left(Y \mid \bar{X}_{t_0}=\bar{x}_{t_0}, \bar{W}_{t_0}=\bar{w}_{t_0}\right) \\
& -\mathbb{E}\left(Y \mid \bar{X}_{t_0}=\bar{x}^*_{t_0}, \bar{W}_{t_0}=\bar{w}_{t_0}\right)\Big] \times \mathbb{P}(\bar{W}_{t_0}=\bar{w}_{t_0}), \\
= & \sum_{\bar{w}_{t_0}\in\Omega_{\bar{W}_{t_0}}} \Big[\mathbb{E}\left(Y \mid \bar{X}_{t_0}=\bar{x}_{t_0}, \bar{W}_{t_0}=\bar{w}_{t_0}\right) \\
& -\mathbb{E}\left(Y \mid \bar{X}_{t_0}=\bar{x}^*_{t_0}, \bar{W}_{t_0}=\bar{w}_{t_0}\right)\Big] \times \mathbb{P}(\bar{W}_{t_0}=\bar{w}_{t_0}), \\
= & \sum_{\bar{w}_{t_0}\in\Omega_{\bar{W}_{t_0}}} \Big[\mathbb{E}\left(Y \mid \bar{X}_{t_0}=\bar{x}_{t_0}, \bar{W}_{t_0}=\bar{w}_{t_0}\right) - \mathbb{E}\left(Y \mid \bar{X}_{t_0}=\bar{x}^*_{t_0}, \bar{W}_{t_0}=\bar{w}_{t_0}\right)\Big] \\
& \times \prod_{t=1}^{t_0} \mathbb{P}(W_t=w_t \mid \bar{W}_{t-1}=\bar{w}_{t-1}),
\end{aligned} \tag{3.12}
$$

for any $\bar{x}_{t_0}$ and $\bar{x}^*_{t_0}$ leading to $\mathscr{X}=\boldsymbol{x}$ and $\mathscr{X}=\boldsymbol{x}^*$.

Now, observe that the conditional ignorability condition $(Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} \perp\!\!\!\perp \bar{X}_{t_0} \mid \bar{W}_{t_0})_{LS}$ does not hold, since $(W_t)_{t>1}$ is affected by the time-varying exposure, but the sequential ignorability condition does hold: $(Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} \perp\!\!\!\perp X_1 \mid W_1)_{LS}$ and $(Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} \perp\!\!\!\perp X_t \mid \{\bar{X}_{t-1}, \bar{W}_t\})_{LS}$ for any time $t \in [\![2;t_0]\!]$. Therefore, for any $\bar{x}_{t_0}$ and $\bar{x}^*_{t_0}$ in $\{0,1\}^{t_0}$,

$$
\begin{aligned}
ATE_{LS}\left(\bar{x}_{t_0};\bar{x}^*_{t_0}\right) = & \ \mathbb{E}_L\left(Y^{\bar{X}_{t_0}=\bar{x}_{t_0}} - Y^{\bar{X}_{t_0}=\bar{x}^*_{t_0}}\right), \\
= & \sum_{\bar{w}_{t_0}\in\Omega_{\bar{W}_{t_0}}} \Big[\mathbb{E}\left(Y \mid \bar{X}_{t_0}=\bar{x}_{t_0}, \bar{W}_{t_0}=\bar{w}_{t_0}\right) \\
& \times \prod_{t=1}^{t_0} \mathbb{P}\left(W_t=w_t \mid \bar{W}_{t-1}=\bar{w}_{t-1}, \bar{X}_{t-1}=\bar{x}_{t-1}\right) \\
& -\mathbb{E}\left(Y \mid \bar{X}_{t_0}=\bar{x}^*_{t_0}, \bar{W}_{t_0}=\bar{w}_{t_0}\right) \\
& \times \prod_{t=1}^{t_0} \mathbb{P}\left(W_t=w_t \mid \bar{W}_{t-1}=\bar{w}_{t-1}, \bar{X}_{t-1}=\bar{x}^*_{t-1}\right)\Big],
\end{aligned} \tag{3.13}
$$

By comparing Equation (3.13) with Equation (3.12), it is clear that $ATE_{SV.Conf}\left(\boldsymbol{x};\boldsymbol{x}^*\right)$ cannot usually be expressed in terms of longitudinal total effects. A noteworthy exception is when $(X_t)_t$ does not affect $(W_t)_t$, that is, when $(W_t)_t$ is a pure confounder: in this case Equation (3.13) coincide with Equation (3.12) and $ATE_{SV.Conf}\left(\boldsymbol{x};\boldsymbol{x}^*\right) \fallingdotseq ATE_{LS}\left(\bar{x}_{t_0};\bar{x}^*_{t_0}\right)$, for any $\bar{x}_{t_0}$ and $\bar{x}^*_{t_0}$ leading respectively to $\mathscr{X}=\boldsymbol{x}$ and $\mathscr{X}=\boldsymbol{x}^*$ (note that, in this case, the versions are irrelevant and condition (*T3.Cond*) holds, so this result also follows from Theorem 4 given in Section 3.4.1 the Main Document).

Now let us turn our attention to the case where model (*SV.Med*) in Figure 3.7 is wrongly considered as the true model. As $(Y^{\mathscr{X}=x} \perp\!\!\!\perp \mathscr{X})_{SV.Med}$, the quantity estimated in practice when working under model (*SV.Med*) would be $ATE_{SV.Med}(\boldsymbol{x};\boldsymbol{x}^*) \fallingdotseq \mathbb{E}\left(Y \mid \mathscr{X}=\boldsymbol{x}\right) - \mathbb{E}\left(Y \mid \mathscr{X}=\boldsymbol{x}^*\right)$. But, we have

$$
\begin{aligned}
ATE_{SV.Med}\left(\boldsymbol{x};\boldsymbol{x}^*\right) \fallingdotseq & \sum_{\bar{x}_{t_0}}\sum_{\bar{x}^*_{t_0}} \left[\mathbb{E}\left(Y \mid \mathscr{X}=\boldsymbol{x}, \bar{X}_{t_0}=\bar{x}_{t_0}\right) - \mathbb{E}\left(Y \mid \mathscr{X}=\boldsymbol{x}^*, \bar{X}_{t_0}=\bar{x}^*_{t_0}\right)\right] \\
& \times \mathbb{P}(\bar{X}_{t_0}=\bar{x}_{t_0} \mid \mathscr{X}=\boldsymbol{x}) \times \mathbb{P}(\bar{X}_{t_0}=\bar{x}^*_{t_0} \mid \mathscr{X}=\boldsymbol{x}^*), \\
= & \sum_{\bar{x}_{t_0}}\sum_{\bar{x}^*_{t_0}} \sum_{\bar{w}_{t_0}\in\Omega_{\bar{W}_{t_0}}} \left[\mathbb{E}\left(Y \mid \mathscr{X}=\boldsymbol{x}, \bar{X}_{t_0}=\bar{x}_{t_0}, \bar{W}_{t_0}=\bar{w}_{t_0}\right)\right. \\
& \times \mathbb{P}(\bar{W}_{t_0}=\bar{w}_{t_0} \mid \bar{X}_{t_0}=\bar{x}_{t_0}) \\
& \quad - \mathbb{E}\left(Y \mid \mathscr{X}=\boldsymbol{x}^*, \bar{X}_{t_0}=\bar{x}^*_{t_0}, \bar{W}_{t_0}=\bar{w}_{t_0}\right) \\
& \quad \left. \times \mathbb{P}(\bar{W}_{t_0}=\bar{w}_{t_0} \mid \bar{X}_{t_0}=\bar{x}^*_{t_0})\right] \\
& \quad \times \mathbb{P}(\bar{X}_{t_0}=\bar{x}_{t_0} \mid \mathscr{X}=\boldsymbol{x}) \times \mathbb{P}(\bar{X}_{t_0}=\bar{x}^*_{t_0} \mid \mathscr{X}=\boldsymbol{x}^*), \\
= & \sum_{\bar{x}_{t_0}}\sum_{\bar{x}^*_{t_0}} \sum_{\bar{w}_{t_0}\in\Omega_{\bar{W}_{t_0}}} \left[\mathbb{E}\left(Y \mid \mathscr{X}=\boldsymbol{x}, \mathscr{W}=\boldsymbol{w}, \bar{X}_{t_0}=\bar{x}_{t_0}, \bar{W}_{t_0}=\bar{w}_{t_0}\right)\right. \\
& \times \mathbb{P}(\bar{W}_{t_0}=\bar{w}_{t_0} \mid \bar{X}_{t_0}=\bar{x}_{t_0}) \\
& \quad - \mathbb{E}\left(Y \mid \mathscr{X}=\boldsymbol{x}^*, \mathscr{W}=\boldsymbol{w}, \bar{X}_{t_0}=\bar{x}^*_{t_0}, \bar{W}_{t_0}=\bar{w}_{t_0}\right) \\
& \quad \left. \times \mathbb{P}(\bar{W}_{t_0}=\bar{w}_{t_0} \mid \bar{X}_{t_0}=\bar{x}^*_{t_0})\right] \\
& \quad \times \mathbb{P}(\bar{X}_{t_0}=\bar{x}_{t_0} \mid \mathscr{X}=\boldsymbol{x}) \times \mathbb{P}(\bar{X}_{t_0}=\bar{x}^*_{t_0} \mid \mathscr{X}=\boldsymbol{x}^*), \\
= & \sum_{\bar{x}_{t_0}}\sum_{\bar{x}^*_{t_0}} \sum_{\bar{w}_{t_0}\in\Omega_{\bar{W}_{t_0}}} \left[\mathbb{E}\left(Y \mid \mathscr{X}=\boldsymbol{x}, \mathscr{W}=\boldsymbol{w}, \bar{X}_{t_0}=\bar{x}_{t_0}, \bar{W}_{t_0}=\bar{w}_{t_0}\right)\right. \\
& \times \prod_{t=1}^{t_0} \mathbb{P}(W_t=w_t \mid \bar{W}_{t-1}=\bar{w}_{t-1}, \bar{X}_{t_0}=\bar{x}_{t_0}) \\
& \quad - \mathbb{E}\left(Y \mid \mathscr{X}=\boldsymbol{x}^*, \mathscr{W}=\boldsymbol{w}, \bar{X}_{t_0}=\bar{x}^*_{t_0}, \bar{W}_{t_0}=\bar{w}_{t_0}\right) \\
& \quad \left. \times \prod_{t=1}^{t_0} \mathbb{P}(W_t=w_t \mid \bar{W}_{t-1}=\bar{w}_{t-1}, \bar{X}_{t_0}=\bar{x}^*_{t_0})\right] \\
& \quad \times \mathbb{P}(\bar{X}_{t_0}=\bar{x}_{t_0} \mid \mathscr{X}=\boldsymbol{x}) \times \mathbb{P}(\bar{X}_{t_0}=\bar{x}^*_{t_0} \mid \mathscr{X}=\boldsymbol{x}^*),
\end{aligned}
$$

with $\boldsymbol{w}$ the value taken by $\mathscr{W}$ when $\bar{W}_{t_0}=\bar{w}_{t_0}$.

On the other hand from Equation (3.13) it follows that

$$
\begin{aligned}
ATE_{LS}\left(\bar{x}_{t_0};\bar{x}^*_{t_0}\right) = & \sum_{\bar{w}_{t_0}\in\Omega_{\bar{W}_{t_0}}} \left[\mathbb{E}\left(Y \mid \mathscr{X}=\boldsymbol{x}, \mathscr{W}=\boldsymbol{w}, \bar{X}_{t_0}=\bar{x}_{t_0}, \bar{W}_{t_0}=\bar{w}_{t_0}\right)\right. \\
& \quad \times \prod_{t=1}^{t_0} \mathbb{P}\left(W_t=w_t \mid \bar{W}_{t-1}=\bar{w}_{t-1}, \bar{X}_{t-1}=\bar{x}_{t-1}\right) \\
& - \mathbb{E}\left(Y \mid \mathscr{X}=\boldsymbol{x}^*, \mathscr{W}=\boldsymbol{w}, \bar{X}_{t_0}=\bar{x}^*_{t_0}, \bar{W}_{t_0}=\bar{w}_{t_0}\right) \\
& \quad \left. \times \prod_{t=1}^{t_0} \mathbb{P}\left(W_t=w_t \mid \bar{W}_{t-1}=\bar{w}_{t-1}, \bar{X}_{t-1}=\bar{x}^*_{t-1}\right)\right],
\end{aligned}
$$

with $\boldsymbol{x}$ and $\boldsymbol{x}^*$ the values taken by $\mathscr{X}$ when $\bar{X}_{t_0}=\bar{x}_{t_0}$ and $\bar{X}_{t_0}=\bar{x}^*_{t_0}$, respectively, and $\boldsymbol{w}$ the value taken by $\mathscr{W}$ when $\bar{W}_{t_0}=\bar{w}_{t_0}$.

By comparing the last two equations, it is clear that $ATE_{SV.Med}\left(\boldsymbol{x};\boldsymbol{x}^*\right)$ cannot usually be expressed in terms of longitudinal total effects. A noteworthy exception is when $(W_t)_t$ does not affect $(X_t)_t$, that is when $(W_t)_t$ is a pure mediator. In this case, $W_t \perp\!\!\!\perp \underline{X}^t_{t_0} \mid \bar{X}_{t-1}$

for any $t \in [\![1; t_0]\!]$, with $\underline{X}_t^{t_0} := (X_t, X_{t+1}, \ldots, X_{t_0})$ which denotes the exposure profile from time t to time t_0, and $ATE_{SV.Med}(\boldsymbol{x}; \boldsymbol{x}^*) \Bumpeq \sum_{\bar{x}_{t_0}} \sum_{\bar{x}_{t_0}^*} ATE_{LS}\left(\bar{x}_{t_0}; \bar{x}_{t_0}^*\right) \times \mathbb{P}(\bar{X}_{t_0} = \bar{x}_{t_0} \mid \mathscr{X} = \boldsymbol{x}) \times \mathbb{P}(\bar{X}_{t_0} = \bar{x}_{t_0}^* \mid \mathscr{X} = \boldsymbol{x}^*)$. Therefore, in this case, $ATE_{SV.Med}(\boldsymbol{x}; \boldsymbol{x}^*)$ coincides with the weighted average given in Equation (3.8) in Section 3.4.1 in the Main Document (note that, in this case, condition (*T3.Uncond*) holds, and this result also follows from Theorem 3 given in Section 3.4.1 in the Main Document).

Chapter 4

On some limitations of probabilistic models for dimension-reduction: illustration in the case of one particular probabilistic formulation of PLS

This Chapter corresponds to the preprint available at https://arxiv.org/abs/2005.09498, and written with Vivian Viallon.

In this Chapter, we are following the notations used by el Bouhaddani et al. (2018); they slightly differ from the ones used in the Introduction or Discussion chapters of this manuscript.

Abstract

Partial Least Squares (PLS) refer to a class of dimension-reduction techniques aiming at the identification of two sets of components with maximal covariance, in order to model the relationship between two sets of observed variables $x \in \mathbb{R}^p$ and $y \in \mathbb{R}^q$, with $p \geq 1, q \geq 1$. el Bouhaddani et al. (2018) have recently proposed a probabilistic formulation of PLS. Under the constraints they consider for the parameters of their model, this latter can be seen as a probabilistic formulation of one version of PLS, namely the PLS-SVD. However, we establish that these constraints are too restrictive as they define a very particular subset of distributions for (x, y) under which, roughly speaking, components with maximal covariance (solutions of PLS-SVD), are also necessarily of respective maximal variances (solutions of the principal components analyses of x and y, respectively). Then, we propose a simple extension of el Bouhaddani et al.'s model, which corresponds to a more general probabilistic formulation of PLS-SVD, and which is no longer restricted to these particular distributions. We present numerical examples to illustrate the limitations of the original model of el Bouhaddani et al. (2018).

4.1 Introduction

Principal Component Analysis (PCA), Canonical Correlation Analysis (CCA) and Partial Least Squares (PLS) are arguably among the most popular multivariate methods for dimension-reduction. They have been described and applied for many years (Hotelling, 1933, 1936, Jöreskog and Wold, 1982, Sampson et al., 1989, Wold, 1985), but are still the subject of active research and discussion (Abdi et al., 2013, Jolliffe, 2002, Jolliffe and Cadima, 2016, Krishnan et al., 2011). Overall, these methods aim at the identification of vectors of weights, from which components are defined as linear transformations of the observed variables. Under each particular method, these weights are chosen so that the corresponding components meet a particular criterion. For example, given a data matrix $\mathbf{X}$ containing n observations of a p-variate variable x (with $n \geq 1$, $p \geq 1$), the goal of PCA is to identify $r \leq p$ unit vectors of weights that define r mutually orthogonal principal components with maximal variances; the matrix of principal components $\mathscr{X} = \mathbf{X}A$ then consists of linear combinations of the p columns of $\mathbf{X}$, with the matrix of weights A given by the eigenvectors associated with the r largest eigenvalues of the sample variance matrix $\mathbf{X}^\top\mathbf{X}$. On the other hand, given two data matrices $\mathbf{X}$ and $\mathbf{Y}$, that gather the $n \geq 1$ observations for a pair of variables (x, y), with $x \in \mathbb{R}^p$ and $y \in \mathbb{R}^q$, $p, q \geq 1$, the goal of CCA and PLS is to model the relationship between x and y by identifying weights that define components with maximal association. Although CCA, which looks for components with maximal correlation, is sometimes considered as a PLS technique, the PLS qualifier usually rather refers to the class of methods that look for components with maximal covariance (Wegelin, 2000). The family of PLS methods still consists of a number of techniques, such as PLS Regression, PLS-W2A or PLS-SVD (Rosipal and Krämer, 2006, Wegelin, 2000). PLS Regression treats the two sets of variables asymmetrically: it focuses on the construction of components from one set of variables, which are then considered as predictors of the second set of variables (the response). On the other hand, both PLS-W2A and PLS-SVD adopt a more symmetrical perspective, and aim at the identification of two sets of weight vectors defining two sets of components. In particular, PLS-SVD, sometimes also referred to as PLS-SB or PLS-C (Krishnan et al., 2011, Sampson et al., 1989, Wegelin, 2000), is simply based on the Singular Value Decomposition (SVD) of the sample covariance matrix $\mathbf{X}^\top\mathbf{Y}$, and defines the two sets of weights as left and right singular vectors of $\mathbf{X}^\top\mathbf{Y}$, respectively. For the sake of completeness, we shall recall that, in contrast, both PLS Regression and PLS-W2A are iterative methods, based on a principle called deflation, which is applied iteratively to guarantee some particular orthogonality properties (Höskuldsson, 1988, Rosipal and Krämer, 2006, Wegelin, 2000, Wold, 1985).

Over the last two decades, several probabilistic formulations of these various dimension-reduction techniques have been introduced, first under a Gaussian setting. They include the Probabilistic PCA (PPCA) (Tipping and Bishop, 1999), the Probabilistic CCA (PCCA) (Bach and Jordan, 2005), as well as several versions of Probabilistic PLS (PPLS)

(el Bouhaddani et al., 2018, Li et al., 2015, Zheng et al., 2016). Regarding these three probabilistic formulations of the PLS, both Zheng et al. (2016) and Li et al. (2015) focus on PLS Regression (the model considered by Li et al. (2015) has commonalities with a probabilistic formulation of Principal Component Regression (PCR) models (Ge et al., 2011)), while el Bouhaddani et al. (2018) consider a symmetrical PLS approach. Overall, all these probabilistic formulations rely on structural equations that define the observed variables as linear combinations of some latent variables plus some Gaussian noise. Parameter estimation under these latent variable models is then usually performed via an Expectation-Maximization (EM) algorithm (Dempster et al., 1977). Giving access to all the likelihood-based inference machinery, these probabilistic formulations have a number of advantages compared to their standard formulation counterpart (Rosipal and Krämer, 2006, Smilde et al., 2004). The estimation can deal with missing data, while still being computationally efficient (Tipping and Bishop, 1999, Zheng et al., 2016). Moreover, covariates can be included in the model (Chiquet et al., 2017), and penalized versions of the likelihood can be used to encourage sparsity or structured sparsity, in particular in a high-dimensional framework (Guan and Dy, 2009, Park et al., 2017, Zeng et al., 2017). Finally, the probabilistic formulation is very versatile, and turns several complex settings into natural extensions of the simple Gaussian ones mentioned above. For example, probabilistic PCA models have been proposed for binary data and count data (Chiquet et al., 2017, Durif et al., 2019). Extensions to even more complex settings, including mediation analysis where three sets of observed variables are involved, have also been proposed (Derkach et al., 2019).

To recap, probabilistic formulations of dimension-reduction techniques enjoy a number of appealing properties. However, appearances can be deceptive, and we will show in this article that some caution is needed when developing and applying them. Indeed, despite their apparent ability to fully capture the relationships among the variables under study, some of them manage to do so under very particular distributions only: when constraints on the model parameters are too strong, the parameters of interest reduce to parameters that could be obtained under much simpler models, which greatly limits the applicability and interest of the corresponding models. For illustration, we will focus here on the probabilistic PLS model proposed by el Bouhaddani et al. (2018), which we will simply refer to as the PPLS model from now on. In Section 4.2.1, we recall the principle of the PPLS model as proposed by el Bouhaddani et al. (2018), and emphasize that it can be regarded as a probabilistic formulation of PLS-SVD. In Section 4.2.2 we show that this PPLS model suffers from the aforementioned defect, and actually defines a set of very particular distributions for (x, y), which limits its applicability. We propose a more general probabilistic formulation of PLS-SVD in Section 4.2.3. In Section 4.3, we present numerical examples to illustrate the limitations of the original PPLS model of el Bouhaddani et al. (2018). Concluding remarks are finally presented in Section 4.4.

4.2 PPLS models

4.2.1 The original PPLS model proposed by el Bouhaddani et al. (2018)

(a) (b)

Figure 4.1: Graphical models for: (a) - The original PPLS model proposed by el Bouhaddani et al. (2018) and recalled in Equation (4.1). (b) - Our extended PPLS model given in Equation (4.3). Note that apart from the fact that the later only has one set of latent variables t, the structure of the noise parts e and f differs between the two models. In both models, x and y are the observed variables whereas circled nodes denote unobserved variables.

The PPLS model proposed by el Bouhaddani et al. (2018) can be graphically represented as depicted in Figure 4.1 (a). More precisely, it is defined by the following structural equations, which relate the two observed sets of variables $x \in \mathbb{R}^p$ and $y \in \mathbb{R}^q$ to two sets of latent variables $t \in \mathbb{R}^r$ and $u \in \mathbb{R}^r$, with $r < \min(p, q)$,

$$x = tW^\top + e, \quad y = uC^\top + f, \quad u = tB + h. \tag{4.1}$$

el Bouhaddani et al. (2018) imposed the constraints (a)-(i) below on the model parameters to ensure identifiability.

(a) $t \sim \mathcal{N}(0, \Sigma_t)$.

(b) Σ_t is a $r \times r$ diagonal matrix, with strictly positive diagonal elements.

(c) $e \sim \mathcal{N}(0_p, \sigma_e^2 I_p)$. (d) $f \sim \mathcal{N}(0_q, \sigma_f^2 I_q)$. (e) $h \sim \mathcal{N}(0_r, \sigma_h^2 I_r)$.

(f) t, e, f and h are independent. u, e and f are independent.

(g) W and C are respectively $p \times r$ and $q \times r$ semi-orthogonal matrices.

(h) B is a diagonal matrix, with strictly positive diagonal elements.

(i) the diagonal elements of $\Sigma_t B$ are strictly decreasingly ordered.

(j) $r < \min(p, q)$.

Here I_p denote the identity matrix of size $p \times p$, and 0_p the vector $(0, \ldots, 0)$ of size p. The parameters of the model are given by $\theta = (W, C, B, \Sigma_t, \sigma_e^2, \sigma_f^2, \sigma_h^2)$. In particular, matrices $W = (W_1, \cdots, W_r)$ and $C = (C_1, \cdots, C_r)$ contain the two sets of weight vectors;

note that they are the "true weights", defined from the theoretical distribution of (x, y). Given estimates $\widehat{W}$ and $\widehat{C}$ of these quantities, two sets of empirical components can be defined as linear combination of the two sets observed variables. In this work, we will mostly focus on components defined as $\widehat{\mathscr{x}} = \mathbf{X}\widehat{W}$ and $\widehat{\mathscr{y}} = \mathbf{Y}\widehat{C}$; we recall that, when working with latent variable models, an alternative strategy consists in using appropriate conditional expectations of the latent variables; see Bach and Jordan (2005) and Section 4.2.3 below for more details. We shall further stress that in either case, the components do not directly correspond to the latent variables t and u. In particular, $\mathscr{x} = xW = t + eW$ and $\mathscr{y} = yC = u + fC$ typically differ from t and u, respectively.

Under the constraints (a)-(j), el Bouhaddani et al. (2018) establish the identifiability of their model (up to sign for the columns of parameters W and C). In particular, the identifiability of parameters W and C is given by the following Proposition.

Proposition 1. *Under the PPLS model given in Equation* (4.1) *along with the constraints* (a)-(j)*, the columns of W and C are the uniquely defined (up to sign) left and right singular vectors corresponding to the r largest singular values of* $\mathrm{Cov}(x, y)$*, respectively.*

This result has already been established by el Bouhaddani et al. (2018) in their Lemma 1, so we here only recall the sketch of the proof. Under the PPLS model, we have $\mathrm{Cov}(x, y) = W\Sigma_t BC^\top$, where W and C are semi-orthogonal matrices, and $\Sigma_t B$ is diagonal with strictly positive decreasingly ordered diagonal elements. It follows that the first r non-null singular values of $\mathrm{Cov}(x, y)$ are all distinct, and are given by the diagonal of $\Sigma_t B$. As a result, the columns of W and C are uniquely defined (up to sign) as the first r left and right singular vectors of $\mathrm{Cov}(x, y)$, respectively.

Although they do not mention it, their model can therefore be regarded as a probabilistic formulation of PLS-SVD. In particular, this means that the two sets of components $\mathscr{x} = xW$ and $\mathscr{y} = yC$ coincide with the two sets of components with maximal covariance, targeted by the PLS-SVD.

However, we establish in Section 4.2.2 that the two sets of weights W and C, which are the theoretical solutions of the PPLS model, are also necessarily the theoretical solutions of two PPCA models for x and y, respectively. In other words, we will see that the PPLS model defines a set of very particular distributions for (x, y) under which the two sets of components with maximal covariance, $\mathscr{x} = xW$ and $\mathscr{y} = yC$, are also necessarily of respective maximal variances.

4.2.2 Limitation of the original PPLS model

Under the PPLS model of el Bouhaddani et al. (2018), the following Proposition, whose proof is given in Appendix 4.A, also holds.

Proposition 2. *Under the PPLS model given in Equation* (4.1) *along with the constraints* (a)-(j)*, the columns of* W *and* C *are eigenvectors corresponding to the* r *largest eigenvalues of* $\mathrm{Var}(x)$ *and* $\mathrm{Var}(y)$*, respectively.*

Proposition 2 notably implies that, under the PPLS model, the two sets of components $\boldsymbol{x} = xW$ and $\boldsymbol{y} = yC$ are not only of maximal covariance (as implied by Proposition 1), but they are also necessarily of respective maximal variances. Equivalently, this comes from the fact that solutions W and C of the PPLS model are also necessarily solutions of two PPCA models, for x and y respectively. More precisely, the PPLS model implies that both x and y fulfill the following PPCA model, presented here for a generic observed variable $z \in \mathbb{R}^d$

$$z = vV^\top + g, \tag{4.2}$$

under the constraints

(α) $v \sim \mathcal{N}(0, \Sigma_V)$.

(β) Σ_V is a $r \times r$ diagonal matrix, with strictly positive diagonal elements.

(γ) $g \sim \mathcal{N}(0_d, \sigma_g^2 I_d)$.

(δ) V is a $d \times r$ semi-orthogonal matrix.

(ϵ) $r < d$.

This PPCA model is a variation of the one introduced by Tipping and Bishop (1999); see Appendix 4.B for more details. First consider this PPCA model for the observed variable $x \in \mathbb{R}^p$. By comparing, on the one hand, constraints (a), (b), (c), (g) and (j) with constraints $(\alpha) - (\epsilon)$, and, on the other hand, Equation (4.2) and the first equation in Equation (4.1), it appears that the unique solution W of the PPLS model necessarily corresponds to one of the possibly many solutions V of this PPCA model for x. More precisely, when the solution of the PPCA model for x is unique (up to sign), that is when the diagonal elements of Σ_t are all distinct, then the r largest eigenvalues of $\mathrm{Var}(x)$ are all of algebraic multiplicity equal to one, the associated eigenvectors are uniquely defined (up to sign), and they correspond to the columns of V. They are also the columns $W_1, \ldots, W_r$ of W, although not necessarily in the same order; columns of W and V are in the same order if, and only if, the diagonal elements of Σ_t are in decreasing order too. Now, if the diagonal elements of Σ_t are not all distinct, then the solution V of the PPCA model for x is not unique, but the columns of W still necessarily constitute one of these solutions, that is one particular set of eigenvectors corresponding to the r largest eigenvalues of $\mathrm{Var}(x)$.

Similarly, the PPLS model implies that the PPCA model above holds for the observed variable $y \in \mathbb{R}^q$ too, and that the unique solution C of the PPLS model necessarily corresponds to one of the possible solutions of this PPCA model for y. More precisely, if

the diagonal elements of $\Sigma_t B^2$ are all distinct, then the columns of C correspond to the uniquely defined r eigenvectors associated with the r largest eigenvalues of $\mathrm{Var}(y)$. On the other hand, if the diagonal elements of $\Sigma_t B^2$ are not all distinct, then the columns of C still constitute one of the solutions of the PPCA model for y; in particular, they are one of the possible sets of eigenvectors for the r largest eigenvalues of $\mathrm{Var}(y)$.

Putting all this together, the PPLS model of el Bouhaddani et al. (2018) corresponds to a model where two PPCA models, one for x and one for y, are related to each other via the third equation in Equation (4.1). Therefore, the weight matrices W and C, solutions of their PPLS model, are also necessarily solutions of two PPCA models for x and y, so that their model defines a subset of very particular distributions for (x, y), under which components $\boldsymbol{x} = xW$ and $\boldsymbol{y} = yC$ are not only of maximal covariance, but also of respective maximal variances. In particular, if the diagonal elements of Σ_t are all distinct, and if the same holds true for $\Sigma_t B^2$, the "solutions" of the two distinct PPCA models are uniquely defined, and then each of the two marginal distributions of x and y are sufficient to respectively identify each of the two sets of weights that define components with maximal covariance. As will be confirmed in Section 4.3, this greatly limits its applicability.

4.2.3 A more general probabilistic formulation of the PLS-SVD

We now present a generalization of the PPLS model of el Bouhaddani et al. (2018), which corrects its main defect and defines a broader set of distributions for (x, y). Our general idea was to keep the same general form as that of el Bouhaddani et al. (2018), but with weaker constraints, in such a way that the weights W and C cannot generally be identified from the marginal distributions of x and y only.

In the PPLS model, assumptions (a)-(j) are related to various aspects of the model: the distributions of the errors terms, the distributions of the latent variables, as well as "direct" constraints on the model parameters $\theta = (W, C, B, \Sigma_t, \sigma_e^2, \sigma_f^2, \sigma_h^2)$. In order to keep the link with the PLS-SVD for our "extended" PPLS model, we still assume that the weights matrices W and C are semi-orthogonal, and that the variance matrices of the latent variables are diagonal. As a start, we thus only relax the constraints (c) and (d) on the isotropy of the variance matrices for the error terms e and f. To be as general as possible, we simply assume that these variance matrices are positive semi-definite, that is that the error terms e and f are two non-degenerate Gaussian vectors. We will therefore replace constraints (c) and (d) by constraints (c*) and (d*) presented below. But then, to preserve the identifiability of the model (see below), we have to consider a model with only one set of latent variables, in the same vein as the PCCA model of Bach and Jordan (2005). Our extended PPLS model, depicted in Figure 4.1 (*b*), is then defined by the

following two structural equations

$$x = tW^\top + e, \quad y = tC^\top + f, \tag{4.3}$$

under the constraints (a), (b), (g), (j) and:

(c*) $e \sim \mathcal{N}(0_p, \Psi_e)$, with Ψ_e a $p \times p$ semi-positive definite matrix.

(d*) $f \sim \mathcal{N}(0_q, \Psi_f)$, with Ψ_f a $q \times q$ semi-positive definite matrix.

(f*) t, e and f are independent.

(h*) the diagonal elements of Σ_t are strictly decreasingly ordered.

Conditions (f*) and (h*) are the analogues of conditions (f) and (h), respectively, in the case where only one set of latent variables is considered. Further observe that $\mathrm{Cov}(x, y) = W\Sigma_t C^\top$, $\mathrm{Var}(x) = W\Sigma_t W^\top + \Psi_e$, and $\mathrm{Var}(y) = C\Sigma_t C^\top + \Psi_f$, where $\theta = (W, C, \Sigma_t, \Psi_e, \Psi_f)$ are the parameters of our model.

We now present the sketch of the proof of the identifiability of our extended PPLS model, which is an adaptation of the one developed by el Bouhaddani et al. (2018); we refer to Appendix 4.C for a more detailed on the proof. Consider two pairs of random variables, (x, y) and $(\widetilde{x}, \widetilde{y})$, drawn from two extended PPLS models, with respective parameters $\theta = (W, C, \Sigma_t, \Psi_e, \Psi_f)$ and $\widetilde{\theta} = (\widetilde{W}, \widetilde{C}, \widetilde{\Sigma}_t, \widetilde{\Psi}_e, \widetilde{\Psi}_f)$, and respective variance-covariance matrices Σ and $\widetilde{\Sigma}$. Now, assume that $\Sigma = \widetilde{\Sigma}$. This is equivalent to

$$\begin{aligned} W\Sigma_t W^\top + \Psi_e &= \widetilde{W}\widetilde{\Sigma}_t\widetilde{W}^\top + \widetilde{\Psi}_e, & (4.4)\\ C\Sigma_t C^\top + \Psi_f &= \widetilde{C}\widetilde{\Sigma}_t\widetilde{C}^\top + \widetilde{\Psi}_f, & (4.5)\\ W\Sigma_t C^\top &= \widetilde{W}\widetilde{\Sigma}_t\widetilde{C}^\top. & (4.6) \end{aligned}$$

Matrices W, C, $\widetilde{W}$, and $\widetilde{C}$ are all semi-orthogonal, and both Σ_t and $\widetilde{\Sigma}_t$ are diagonal with strictly decreasing diagonal elements. As detailed in Appendix 4.C, Equation (4.6) implies that $\Sigma_t = \widetilde{\Sigma}_t$, $W = \widetilde{W}J$ and $C = \widetilde{C}J$, with J a diagonal matrix with ± 1 elements on the diagonal. Then, Equation (4.4) implies that $\Psi_e = \widetilde{\Psi}_e$, while Equation (4.5) implies that $\Psi_f = \widetilde{\Psi}_f$. As a result, the parameters of the extended PPLS model given in Equation (4.3) are identifiable (up to sign for the columns of W and C). In particular, because $\mathrm{Cov}(x, y) = W\Sigma_t C^\top$, parameters W and C are identified (up to sign) as the first r left and right singular vectors of $\mathrm{Cov}(x, y)$, respectively.

Moreover, because $\mathrm{Var}(x) = W\Sigma_t W^\top + \Psi_e$, and $\mathrm{Var}(y) = C\Sigma_t C^\top + \Psi_f$, with Ψ_e and Ψ_f two positive semi-definite matrices, we shall stress that W and C can generally not be identified from the eigendecomposition of $\mathrm{Var}(x)$ and $\mathrm{Var}(y)$, respectively. In other words, the two sets of weights W and C define components with maximal covariance, which are not necessarily of respective maximal variances, and W and C cannot generally

be identified separately from the marginal distributions of x and y. Our extended PPLS model can therefore be regarded as a more general probabilistic formulation of the PLS-SVD, which defines a much broader and interesting set of distributions than the original PPLS model of el Bouhaddani et al. (2018).

We will now conclude this Section by a few remarks on our model. First, we shall stress that the residuals, e and f of our model, may be more than simple noise terms. Indeed, they consist of everything that is not in the shared part between x and y. In particular, e may contain some signal from additional latent variables specific to x, plus some pure noise. Similarly, f may contain some signal from additional latent variables specific to y.

Second, two sets of components can be defined as linear transformations of x and y, respectively. As above, just as under the standard PLS-SVD (Wegelin, 2000), a first strategy consists in defining $\boldsymbol{x} = xW$ and $\boldsymbol{y} = yC$. Following Bach and Jordan (2005), alternative components are defined as $\boldsymbol{x}^* = \mathrm{E}(t|x;\theta)$ and $\boldsymbol{y}^* = \mathrm{E}(t|y;\theta)$. As $\mathrm{E}(t|x;\theta) = x\big(W\Sigma_t W^\top + \Psi_e\big)^{-1} W\Sigma_t$ and $\mathrm{E}(t|y;\theta) = y\big(C\Sigma_t C^\top + \Psi_f\big)^{-1} C\Sigma_t$, these components are linear transformations of x and y too, but yield different linear sub-spaces than $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively, unless Ψ_e and Ψ_f are zero matrices (Bach and Jordan, 2005).

Finally, a last remark concerns the estimation of the parameters under our model. Parameters W and C could be estimated by performing a simple SVD of the covariance matrix $\mathrm{Cov}(\mathbf{X},\mathbf{Y})$. Alternatively, an EM-algorithm would yield estimates for all the parameters θ, while taking into account all the constraints of the model. It would further allow various extensions, such as the inclusion of covariates, etc. However, the derivation of the EM algorithm is less straightforward under our extended model than under the original PPLS model. In particular, the updates in each of the M-steps of the EM for the parameters W and C require an optimization problem over the Stiefel Manifold to be solved (Siegel, 2019, Wen and Yin, 2010), while these updates have closed form expressions under the original PPLS model of el Bouhaddani et al. (2018). Although we have not fully devised it, additional details on a possible EM algorithm are presented in Appendix 4.D.

4.3 Simulation study

Now, we present results from two simulation studies aimed to illustrate the limitations of the original PPLS model, and, more precisely, to illustrate the behavior of the estimates for W and C returned by the EM algorithm devised by el Bouhaddani et al. (2018) under the original PPLS model, depending on whether this model is correctly specified or not. For comparison, we further considered estimates returned by the standard (non-probabilistic) PLS-SVD, and the standard PCA (successively applied on the "x and y parts" of the data). The PLS-W2A, which is another symmetrical PLS method that we briefly described in

the Introduction (see Rosipal and Krämer (2006), Wegelin (2000), Wold (1985) for more details), was originally considered too. As expected, estimates returned by the PLS-W2A and PLS-SVD methods were very similar under the original PPLS model (because $\text{Var}(xW)$ and $\text{Var}(yC)$ are diagonal under the original PPLS model), but as they were in the second simulation study too, we finally decided to omit their presentation here.

We set the dimensions of the observed sets of variables x and y to $p = q = 20$, the dimension of the sets of latent variables to $r = 3$, and make the sample size vary in $n \in \{50, 250, 500, 1000, 5000\}$. In the first simulation study, we work under the same setting as that considered by el Bouhaddani et al. (2018) in their simulation study. More precisely, data $(\mathbf{X}, \mathbf{Y})$ are generated under the original PPLS model, in the particular case where the diagonal elements of both Σ_t and $\Sigma_t B^2$ are all distinct. Weight matrices W and C are randomly drawn from the sets of semi-orthogonal matrices of size $p \times r$ and size $q \times r$, respectively, and the diagonal elements of Σ_t and B are respectively set to $\sigma^2_{t_i} = \exp(-(i-1)/5)$ and $b_i = 1.5\exp(3(i-1)/10)$, for $i \in \{1, 2, 3\}$, just as el Bouhaddani et al. (2018). As for the variances of e, f and h, they are chosen so that the signal-to-noise ratios are equal to 0.25: $\sigma^2_e = 0.4$, $\sigma^2_f = 4$ and $\sigma^2_h = 5.33$. The main objective of this first study is to empirically confirm that, when the original PPLS model of el Bouhaddani et al. (2018) is correctly specified, the weights returned by the corresponding EM algorithm are similar to those returned by two PCAs applied on the x and y parts of the data. In the second simulation study, data are generated under a model similar to the original PPLS model, except that e and f are not of isotropic variance; instead e and f are drawn from multivariate Gaussian variables with arbitrary positive semi-definite variance matrices; more precisely, we chose positive-definite matrices ensuring that eigenvectors of matrices $\text{Var}(x)$ and $\text{Var}(y)$ were not too close to the left and right singular vectors of $\text{Cov}(x, y)$ (using a simple acceptance rejection method), to make sure we work under really misspecified models where solutions of the PLS-SVD differ from solutions of two PCAs. The main objective of this second study is to describe how the solutions of the EM algorithm of el Bouhaddani et al. (2018) behaves when components of maximal covariance are not of respective maximal variances too, that is when the original PPLS model is misspecified. In both studies, the results are computed over 1000 replicates. For the comparisons of weight vectors, we use the cosine similarity, which simply reduces to the dot product in our case since the true and estimated weight vectors are unit vectors. Results from our simulation studies can be replicated using our R scripts that we will make soon available on GitHub.

Figure 4.2 presents the median of the cosine similarity (in absolute values) between the true weights W and C and their estimates, computed under the original PPLS model (first row), and under our extended PPLS model (second row). Each of the three columns of Figure 1 presents the results for one particular pair $(W_i, C_i)_{i \in \{1,2,3\}}$. Following what el Bouhaddani et al. (2018) did in their simulation study, we shall stress that the columns

of the estimated weight matrices returned by each of the three compared methods were first re-arranged to make sure they matched the ordering of the true weight matrices.

When the PPLS model is correctly specified (top panel of Figure 4.2), estimates returned by the EM algorithm under the original PPLS models perform similarly to estimates returned by the other PLS techniques (PLS-SVD and PLS-W2A), and they are all reasonably close to the true weight vectors. In particular, their cosine similarity with the true weight vectors tend to 1 as sample size increases. But, as expected, this is also the case for the estimates returned by two PCAs successively applied on $\mathbf{X}$ and $\mathbf{Y}$. This empirically confirms that when the diagonal elements of both Σ_t and $\Sigma_t B^2$ are all distinct under the original PPLS model, solutions of the PLS-SVD coincide with those of the PCAs (keep in mind that when the diagonal elements of Σ_t and/or $\Sigma_t B^2$ are not all distinct, solutions of the PLS-SVD still constitute one of the solutions of the PCAs).

On the other hand, when the original PPLS model is misspecified (bottom panel of Figure 4.2), our results show that, estimates returned by the two PCAs are quite far from the true weight vectors (as expected, by design), while those returned by the PLS-SVD still perform well. As for the EM algorithm devised under the original PPLS model, it performs much worse than the PLS-SVD, and not much better than the two PCAs. To better describe the estimates returned by the EM algorithm devised under the original PPLS model, Figure 4.3 presents the median of the cosine similarities (in absolute values) between these estimates and those returned by the PLS-SVD and the two distinct PCAs. Interestingly, these results show that, on average, estimates returned by the EM algorithm under the original PPLS model are closer to those returned by the PCAs, especially when the original PPLS model is misspecified. Figure 4.4 in Appendix 4.E further presents the box-plots of the absolute value of the cosine similarities between the estimates returned by the EM algorithm devised under the original PPLS model and those returned by (*i*) two distinct PCAs, and (*ii*) the standard PLS-SVD, in our second simulation study (when the original PPLS model is misspecified). These box-plots suggest that, when solutions of the PLS-SVD differ from solutions of two PCAs, estimates returned by the EM algorithm proposed by el Bouhaddani et al. (2018) are generally closer to those returned by the two PCAs. This constitutes a severe limitation for this algorithm: in real-life examples, there is no guarantee that the estimated weight vectors it returns really capture the relationship between x and y.

4.4 Discussion

In this article, we focused on the PPLS model proposed by el Bouhaddani et al. (2018). After highlighting that it corresponds to a probabilistic formulation of PLS-SVD, we showed that the constraints considered in this original PPLS model are too strong: they imply that the weight matrices W and C, which are solutions of this original PPLS model,

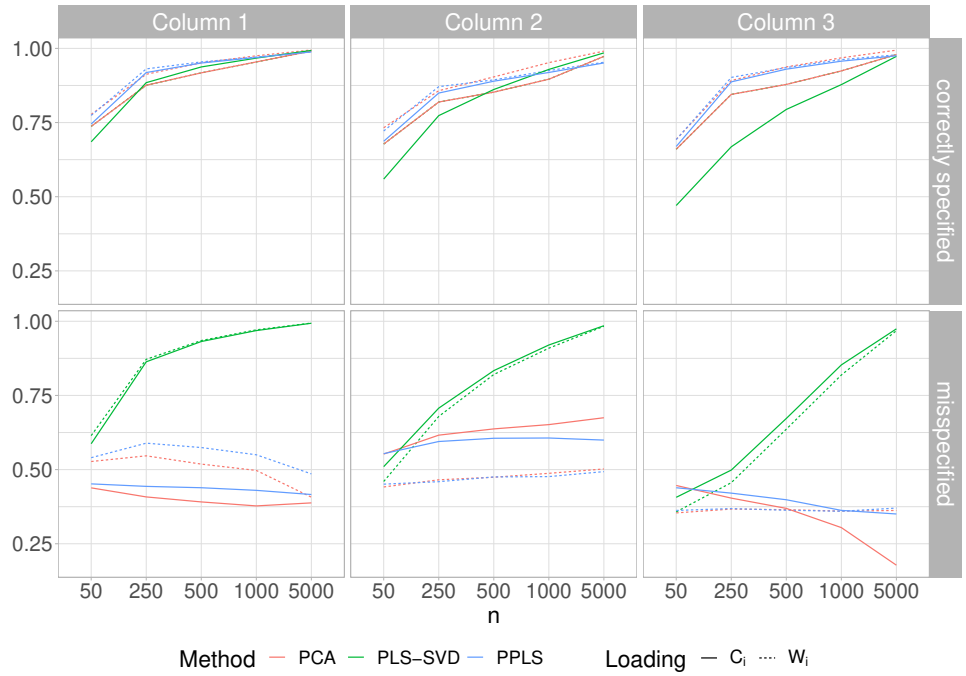


Figure 4.2: Medians of the cosine similarities (in absolute values) between the true weight vectors and the estimates returned by (i) the PPLS EM algorithm, (ii) two distinct PCAs on $\mathbf{X}$ and $\mathbf{Y}$, and (iii) PLS-SVD on $(\mathbf{X}, \mathbf{Y})$. The results are computed over 1000 replicates, for $p = q = 20$, $r = 3$ and different sample sizes $n \in \{50, 250, 500, 1000, 5000\}$. The top panels correspond to the first simulation study where the original PPLS model is correctly specified, while the bottom panels correspond to the second simulation study where the original PPLS model is misspecified.

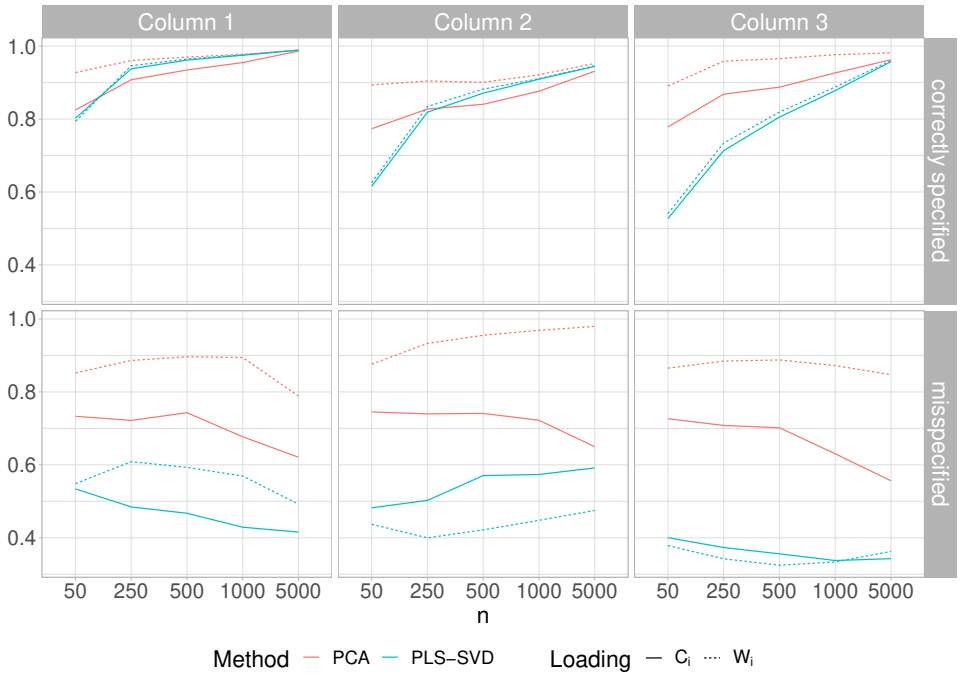


Figure 4.3: Medians of the cosine similarities (in absolute values) between the weight vector estimates returned by the EM algorithm devised under the original PPLS model, and those returned by (i) two distinct PCAs on $\mathbf{X}$ and $\mathbf{Y}$, and (ii) PLS-SVD on $(\mathbf{X}, \mathbf{Y})$. The results are computed over 1000 replicates, for $p = q = 20$, $r = 3$ and different sample sizes $n \in \{50, 250, 500, 1000, 5000\}$. The top panels correspond to the first simulation study where the original PPLS model is correctly specified, while the bottom panels correspond to the second simulation study where it is misspecified.

are also necessarily solutions of two distinct PPCA models for x and y, respectively. As a result, the original PPLS model defines a very particular subset of distributions for the pair (x, y), under which the two sets of components of maximal covariance are necessarily of respective maximal variances too. This defect severely limits the practical interest of this model.

However, this defect might not be specific to the model proposed by el Bouhaddani et al. (2018). Our results more generally stress that some caution is needed when developing and applying such latent variable models for dimension-reduction: when imposing too strong of constraints on the model parameters, a model whose structural equations seem to correctly describe the relationships between the observed variables, may turn out to be too simplistic. It can define very particular distributions, under which parameters of interest could be obtained under much simpler models. As a result, a close inspection of other probabilistic models might be needed. First consider the case of the Probabilistic PLS Regression (PPLS-R) model proposed by Li et al. (2015). Zheng et al. (2016) already suggested this model shared some similarities with the probabilistic formulation of Principal Component Regression (PPCR) proposed by Ge et al. (2011). As a matter of fact, it seems that the weight matrix in the PPLS-R model proposed by Li et al. (2015) could also be defined from the marginal distribution of the predictors only. Similar concerns may apply to more complex frameworks, such as the mediation analysis, where the objective is to describe the relationships between three sets of variables. For example, Derkach et al. (2019) propose an interesting probabilistic formulation, but it might be worth checking whether the joint distribution of the three sets of variables is really needed to identify their parameters of interest, or whether the constraints they considered are also too strong, and define particular distributions under which these parameters can actually be identified using, e.g., the marginal distribution of one particular set of variables.

As shown in the present article, it is sometimes possible to correct for these defects. In the case of the PPLS model originally proposed by el Bouhaddani et al. (2018), we were able to relax some of the constraints, and develop a more general probabilistic formulation of the PLS-SVD, under which the joint distribution of (x, y) is generally necessary for the identification of the model parameters. However, the implementation of an EM algorithm for the estimation of the parameters under this extended PPLS model is less straightforward than for the original PPLS model. In particular, each M-step of the algorithm requires a numerical optimization step to update the estimates of the parameters W and C, whereas such updates are given by closed-form expressions under the original PPLS model. Alternatively, we could propose another version of the model, where parameters W and C would not have to be semi-orthogonal matrices. However, for the model to be identifiable, we would have to impose $\Sigma_t = I_r$ (identifiability would then hold up to an orthogonal transformation for parameters W and C), and the corresponding

model would actually be the PCCA model proposed by Bach and Jordan (2005).

Disclaimer

Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer/World Health Organization.

4.A Appendix: Proof of Proposition 2

Here, we prove that the columns of W and C, solutions of the original PPLS model, are also necessarily eigenvectors corresponding to the r largest eigenvalues of $\mathrm{Var}(x)$ and Var(y), respectively.

Under the PPLS proposed by el Bouhaddani et al. (2018) recalled in Equation (4.1), we have $\mathrm{Var}(x) = W\Sigma_t W^\top + \sigma_e^2 I_p$, with W a semi-orthogonal $p \times r$ matrix, Σ_t a $r \times r$ diagonal matrix and $W\Sigma_t W^\top$ a symmetric $p \times p$ matrix of rank $r < p$. Consider any eigendecomposition $Q\Delta Q^\top$ of matrix $W\Sigma_t W^\top$: Δ is then diagonal, with r non-null elements. Moreover, because W is semi-orthogonal and Σ_t (square) diagonal (with strictly positive diagonal elements), the r non-null eigenvalues in Δ are the diagonal elements $(\sigma_{t_1}^2, \ldots, \sigma_{t_r}^2)$ of Σ_t, and the columns of W are eigenvectors corresponding to these r non-null eigenvalues. Moreover, because Q is orthogonal, we have $\mathrm{Var}(x) = W\Sigma_t W^\top + \sigma_e^2 I_{p_X} = Q\Delta_2 Q^\top$ with Δ_2 the $p \times p$ diagonal matrix with diagonal elements $(\sigma_{t_1}^2 + \sigma_e^2, \ldots, \sigma_{t_r}^2 + \sigma_e^2, \sigma_e^2, \ldots, \sigma_e^2)$. Putting all this together, it follows that the columns of W are eigenvectors of $\mathrm{Var}(x)$ corresponding to its r largest eigenvalues. When the diagonal elements of Σ_t are all distinct, the r non-null eigenvalues of $W\Sigma_t W^\top$ are of algebraic multiplicity equal to one, and so are the r largest eigenvalues of $\mathrm{Var}(x)$. In this case, the columns of W are the uniquely defined r eigenvectors associated with the r largest eigenvalues of $\mathrm{Var}(x)$. However, because $(\sigma_{t_1}^2, \ldots, \sigma_{t_r}^2)$ are not necessarily decreasingly ordered, the eigenvectors of $\mathrm{Var}(x)$ associated with the r largest eigenvalues are not necessarily given in the same order as the left singular vectors associated with the r largest singular values of $\mathrm{Cov}(x, y)$.

The PPLS model of el Bouhaddani et al. (2018) also implies that $\mathrm{Var}(y) = C(\Sigma_t B^2 + \sigma_h^2 I_r)C^\top + \sigma_f^2 I_q$, with C a semi-orthogonal $q \times r$ matrix, and $\Sigma_t B^2 + \sigma_h^2 I_r$ a square diagonal matrix of size $r < q$. Arguing as above, it can be shown that the columns of C constitute one particular set of eigenvectors corresponding to the r largest eigenvalues of $\mathrm{Var}(y)$. In particular, when the diagonal elements of $\Sigma_t B^2$ are distinct, the r largest eigenvalues of $\mathrm{Var}(y)$ are of algebraic multiplicity equal to one: the associated eigenvectors are uniquely defined (up to sign), and they correspond to the columns of C.

4.B Appendix: Additional details for the comparison of the PPCA model given in Equation (4.2) with the one proposed by Tipping and Bishop (1999)

Consider a matrix $\mathbf{Z}$ containing $n \geq 1$ observations of a random variable $z \in \mathbb{R}^d$, $d > 1$. We recall that the principle of the standard PCA applied on $\mathbf{Z}$ is to identify a matrix V or $r \leq d$ vectors of weights defining a matrix $\mathscr{Z} = \mathbf{Z}V$ of r mutually orthogonal components with maximal variances. The matrix of weights V is then simply given by the eigenvectors associated with the r largest eigenvalues of the sample variance matrix $\mathbf{Z}^\top\mathbf{Z}$.

Tipping and Bishop (1999) proposed a probabilistic formulation of PCA, actually inspired by the factor analysis model (Basilevsky, 1994). Their PCCA is defined by the following structural equation

$$z = vA^\top + g, \tag{4.7}$$

under the constraints that $v \sim \mathcal{N}(0_r, I_r)$, $g \sim \mathcal{N}(0_d, \sigma^2 I_d)$ and $r < d$. Then, because $\mathrm{Var}(z)$ is the identity matrix, A is only identifiable up to an orthogonal transformation. In addition, because A is not necessarily semi-orthogonal, the r eigenvectors of $AA^\top$ associated with the r largest eigenvalues have first to be computed to retrieve weights similar to those defined in the non-probabilistic PCA framework.

On the other hand, the PCCA model introduced in Section 4.2.3 is given by the following structural equation

$$z = vV^\top + g, \tag{4.8}$$

along with the constraints $v \sim \mathcal{N}(0_r, \Sigma_v)$, Σ_v is a $r \times r$ diagonal matrix (with strictly positive diagonal elements), $g \sim \mathcal{N}(0_d, \sigma^2 I_d)$, V is a $d \times r$ semi-orthogonal matrix and $r < d$. Under this model, if the diagonal elements of Σ_v are distinct, then the r non-null eigenvalues $\mathrm{Var}(x)$ are distinct, and the columns V are uniquely defined (up to sign) as the associated eigenvectors. On the other hand, if some diagonal elements of Σ_v are identical, the eigenvectors associated with the identical eigenvalues are defined only up to a rotation. In any case, components defined as zV are those (possibly non-uniquely) defined in the non-probabilistic PCA.

Finally, we shall stress that these two models are equivalent, in the following sense. First note that under the setting of Equation (4.7), $AA^\top$ is of size $d \times d$ and of rank $r < d$, and can then always be decomposed either as $Q\Delta Q^\top$, where Δ is a $d \times d$ diagonal matrix with strictly positive r first diagonal elements, and Q is a $d \times d$ orthogonal matrix, or as $Q_r\Delta_r Q_r^\top$, with Δ_r a $r \times r$ diagonal matrix with strictly positive diagonal elements, and Q_r a $d \times r$ semi-orthogonal matrix. Consequently, any solution A of the PPCA model given in Equation (4.7) defines a solution V of the PPCA model given in Equation (4.8), with $V = Q_r$ and $\Sigma_v = \Delta_r$. Similarly, for any solution V of the PPCA model given in Equation (4.8), $A = V\Sigma_v^{\frac{1}{2}}$ is solution of the PPCA model given in Equation (4.7).

4.C Appendix: Proof of the identifiability of the more general probabilistic formulation of PLS-SVD

Our proof is an adaptation of the one presented by el Bouhaddani et al. (2018). Consider two pairs of variables (x, y) and $(\widetilde{x}, \widetilde{y})$ defined under our extended PPLS model given in Equation (4.3), with respective parameters sets $\theta = (W, C, \Sigma_t, \Psi_e, \Psi_f)$ and $\widetilde{\theta} = (\widetilde{W}, \widetilde{C}, \widetilde{\Sigma}_t, \widetilde{\Psi}_e, \widetilde{\Psi}_f)$. Denote by Σ and $\widetilde{\Sigma}$ their variance-covariance matrices. The principle of the proof is to show that $\Sigma = \widetilde{\Sigma}$ implies $\theta = \widetilde{\theta}$. So, let us now assume that $\Sigma = \widetilde{\Sigma}$, that is

$$
\begin{aligned}
W\Sigma_t W^\top + \Psi_e &= \widetilde{W}\widetilde{\Sigma}_t\widetilde{W}^\top + \widetilde{\Psi}_e, && (4.9)\\
C\Sigma_t C^\top + \Psi_f &= \widetilde{C}\widetilde{\Sigma}_t\widetilde{C}^\top + \widetilde{\Psi}_f, && (4.10)\\
W\Sigma_t C^\top &= \widetilde{W}\widetilde{\Sigma}_t\widetilde{C}^\top. && (4.11)
\end{aligned}
$$

We recall that W, C, $\widetilde{W}$, and $\widetilde{C}$ are semi-orthogonal matrices (of respective sizes $p \times r$ and $q \times r$), while Σ_t and $\widetilde{\Sigma}_t$ are diagonal matrices (of size $r \times r$) with strictly decreasingly ordered diagonal elements.

First consider Equation (4.11). $W\Sigma_t C^\top$ is a $p \times q$ matrix of rank r, with $r < \min(p, q)$. Consider any particular singular value decomposition $V\Delta Q^\top$ of matrix $W\Sigma_t C^\top$, with V a square orthogonal matrix of size p, Q a square orthogonal matrix of size q, and Δ a rectangular diagonal matrix of size $p \times q$ with r non-null diagonal elements. The columns of V are eigenvectors of $W\Sigma_t^2 W^\top$, while those of Q are eigenvectors of $C\Sigma_t^2 C^\top$. Moreover, the diagonal elements of Δ are the square roots of the eigenvalues of both $W\Sigma_t^2 W^\top$ and $C\Sigma_t^2 C^\top$. Then, because W (respectively C) is a semi-orthogonal matrix and Σ_t is diagonal, the columns of W (respectively of C) are left (respectively right) singular vectors associated with the r non-null singular values of the matrix $W\Sigma_t C^\top$, which correspond to the diagonal elements of Σ_t. Moreover, since the diagonal elements of Σ_t are strictly decreasingly ordered, these r non-null singular values are distinct, and the r associated left and right singular vectors are uniquely defined (up to sign). Similarly, write $\widetilde{V}\widetilde{\Delta}\widetilde{Q}^\top$ any particular singular value decomposition of $\widetilde{W}\widetilde{\Sigma}_t\widetilde{C}^\top$. From the uniqueness of the singular values and of the first r left and right singular vectors (up to sign), it follows that (i) $\Delta = \widetilde{\Delta}$, (ii) the first r columns of V are equal (up to sign) to the first r columns of $\widetilde{V}$, and (iii) the first r columns of Q are equal (up to sign) to the first r columns of $\widetilde{Q}$. In other words, we have

$$
\begin{aligned}
\Sigma_t &= \widetilde{\Sigma}_t,\\
W &= \widetilde{W}J,\\
C &= \widetilde{C}J,
\end{aligned}
$$

where J is a diagonal matrix with ± 1 diagonal elements. Then, Equation (4.9) is equivalent to $W\Sigma_t W^\top + \Psi_e = W\Sigma_t W^\top + \widetilde{\Psi}_e$, which yields $\Psi_e = \widetilde{\Psi}_e$. In the same way, Equation (4.10) is equivalent to $C\Sigma_t C^\top + \Psi_f = C\Sigma_t C^\top + \widetilde{\Psi}_f$, so that $\Psi_f = \widetilde{\Psi}_f$. As a result, the parameters of our extended PPLS model given in Equation (4.3) are all identifiable (up to sign for the columns of W and C).

4.D Appendix: Details on an EM algorithm for the estimation of the parameters of the PPLS model given in Equation (4.3)

As before, we denote by $(\mathbf{X}, \mathbf{Y}) = \big((X_1, \dots X_n)^\top, (Y_1, \dots Y_n)^\top\big)$ the observed sample of n independent and identically distributed replica of (x, y). On the other hand, we denote by $\mathbf{T} = (T_1, \dots T_n)^\top$ the n "observations" of the latent variable t (which are therefore not observed). To estimate $\theta = (W, C, \Sigma_t, \Psi_e, \Psi_f)$ from $(\mathbf{X}, \mathbf{Y})$, an EM algorithm (Dempster et al., 1977) can be used, as a closed-form for $\widehat{\theta}$ cannot be obtained by directly maximizing the likelihood of the observed data. The main steps of this EM algorithm are briefly described below, especially to highlight the step that requires an optimization on Stiefel Manifolds.

The observed data likelihood is

$$L(\mathbf{X}, \mathbf{Y}; \theta) = \int_{\mathbf{T}} L(\mathbf{X}, \mathbf{Y}, \mathbf{T}; \theta) d\mathbf{T},$$

where the complete-data likelihood $L(\mathbf{X}, \mathbf{Y}, \mathbf{T}; \theta)$ is given by

$$\begin{aligned} L(\mathbf{X}, \mathbf{Y}, \mathbf{T}; \theta) &= \prod_{i=1}^{n} f(X_i, Y_i, T_i; \theta), \\ &= \prod_{i=1}^{n} f_{X_i|T_i}(X_i \mid T_i; \theta)\, f_{Y_i|T_i}(Y_i \mid T_i; \theta)\, f_{T_i}(T_i; \theta), \end{aligned}$$

as $X_i \perp\!\!\!\perp Y_i \mid T_i$, $i \in [\![1, n]\!]$. Under the extended PPLS model, $X_i \mid \{T_i; \theta\} \sim \mathcal{N}\big(T_i W^\top, \Psi_e\big)$, $Y_i \mid \{T_i; \theta\} \sim \mathcal{N}\big(T_i C^\top, \Psi_f\big)$ and $T_i; \theta \sim \mathcal{N}\big(0_r, \Sigma_t\big)$, $i \in [\![1, n]\!]$.

Consequently, for any $i \in [\![1, n]\!]$, we have

$$\big(X_i, Y_i, T_i; \theta\big) \sim \mathcal{N}\left((0_{p+q+r}), \begin{pmatrix} W\Sigma_t W^\top + \Psi_e & W\Sigma_t C^\top & W\Sigma_t \\ C\Sigma_t W^\top & C\Sigma_t C^\top + \Psi_f & C\Sigma_t \\ \Sigma_t W^\top & \Sigma_t C^\top & \Sigma_t \end{pmatrix}\right).$$

Denote by Σ the first $(p+q) \times (p+q)$ block of this variance-covariance matrix. Then, $T_i \mid \{X_i, Y_i\}$ is Gaussian, with $\mathbb{E}\big(T_i \mid X_i, Y_i\big) = \Big(X_i, Y_i\Big)\Sigma^{-1} \begin{pmatrix} W\Sigma_t \\ C\Sigma_t \end{pmatrix}$ and $\mathbb{V}\big(T_i \mid X_i, Y_i\big) =$

$\Sigma_t - \Big(\Sigma_t W^\top, \Sigma_t C^\top\Big)\Sigma^{-1}\begin{pmatrix} W\Sigma_t \\ C\Sigma_t \end{pmatrix}$. Then $\mathbb{E}\big(T_i^\top T_i \mid X_i, Y_i\big) = \mathbb{V}\big(T_i \mid X_i, Y_i\big) + \mathbb{E}\big(T_i \mid X_i, Y_i\big)^\top \mathbb{E}\big(T_i \mid X_i, Y_i\big)$.

For simplicity, we will use the notations

$$\mathbb{E}\big(\mathbf{T}^\top\mathbf{T} \mid \mathbf{X}, \mathbf{Y}; \theta\big) = \sum_{i=1}^{n} \mathbb{E}\big(T_i^\top T_i \mid X_i, Y_i; \theta\big)$$
$$\mathbb{E}\big(\mathbf{T} \mid \mathbf{X}, \mathbf{Y}; \theta\big) = \Big(\mathbb{E}\big(T_i \mid X_i, Y_i; \theta\big)\Big)_{i\in[\![1,n]\!]}$$

which are a square matrix of size r, and a $n \times r$ matrix, respectively.

From any initial value for θ, the EM algorithm consists in successively iterating two steps, namely the E-step and the M-step. From a value θ^{old} for the set of parameters, the conditional moments $\mathbb{E}\big(\mathbf{T}^\top\mathbf{T} \mid \mathbf{X}, \mathbf{Y}, \theta^{old}\big)$ and $\mathbb{E}\big(\mathbf{T} \mid \mathbf{X}, \mathbf{Y}, \theta^{old}\big)$ are computed; this is the E-step. Then the M-step consists in updating the values of the parameters, that is finding θ^{new}, the value of θ which maximizes $\mathbb{E}\Big(\ln\big(L(\mathbf{X}, \mathbf{Y}, \mathbf{T}; \theta)\big) \mid \mathbf{X}, \mathbf{Y}; \theta^{old}\Big)$. In our case in the M-step, we can successively maximize over θ the three following quantities: $\kappa = \mathbb{E}\Big(\ln\big(f_{\mathbf{X}|\mathbf{T}}(\mathbf{X} \mid \mathbf{T}; \theta)\big) \mid \mathbf{X}, \mathbf{Y}; \theta^{old}\Big)$, $\mu = \mathbb{E}\Big(\ln\big(f_{Y|\mathbf{T}}(\mathbf{Y}|\mathbf{T}; \theta)\big) \mid \mathbf{X}, \mathbf{Y}; \theta^{old}\Big)$ and $\pi = \mathbb{E}\Big(\ln\big(f_{\mathbf{T}}(\mathbf{T}; \theta)\big) \mid \mathbf{X}, \mathbf{Y}; \theta^{old}\Big)$, which are here given by

$$\begin{aligned}
\kappa &= -\frac{np}{2}\ln(2\pi) - \frac{n}{2}\ln(|\Psi_e|) - \frac{1}{2}\mathrm{Tr}\Big(\big(\mathbf{X}^\top\mathbf{X} - 2\mathbf{X}^\top\mathbb{E}\big(\mathbf{T} \mid \mathbf{X}, \mathbf{Y}, \theta^{old}\big)W^\top \\
&\qquad + W\mathbb{E}\big(\mathbf{T}^\top\mathbf{T} \mid \mathbf{X}, \mathbf{Y}, \theta^{old}\big)W^\top\big)\Psi_e^{-1}\Big). \\
\mu &= -\frac{nq}{2}\ln(2\pi) - \frac{n}{2}\ln(|\Psi_f|) - \frac{1}{2}\mathrm{Tr}\Big(\big(\mathbf{Y}^\top\mathbf{Y} - 2\mathbf{Y}^\top\mathbb{E}\big(\mathbf{T} \mid \mathbf{X}, \mathbf{Y}, \theta^{old}\big)C^\top \\
&\qquad + C\mathbb{E}\big(\mathbf{T}^\top\mathbf{T} \mid \mathbf{X}, \mathbf{Y}, \theta^{old}\big)C^\top\big)\Psi_f^{-1}\Big). \\
\pi &= -\frac{nr}{2}\ln(2\pi) - \frac{n}{2}\ln\left(|\Sigma_t|\right) - \frac{1}{2}\mathrm{Tr}\Big(\mathbb{E}\big(\mathbf{T}^\top\mathbf{T} \mid \mathbf{X}, \mathbf{Y}, \theta^{old}\big)\Sigma_t^{-1}\Big).
\end{aligned}$$

Of course, the updated parameter θ^{new} has to fulfill the constraints of our model. In particular, solutions W^{new} and C^{new} are defined as the semi-orthogonal matrices W and C that maximize κ and μ above. More precisely,

$$\begin{aligned}
W^{new} &= \underset{W}{\mathrm{argmax}} -\frac{1}{2}\mathrm{Tr}\Big(\big(\mathbf{X}^\top\mathbf{X} - 2\mathbf{X}^\top\mathbb{E}\big(\mathbf{T} \mid \mathbf{X}, \mathbf{Y}, \theta^{old}\big)W^\top + W\mathbb{E}\big(\mathbf{T}^\top\mathbf{T} \mid \\
&\qquad \mathbf{X}, \mathbf{Y}, \theta^{old}\big)W^\top\big)\Psi_e^{old^{-1}}\Big) \qquad \text{s.t.} \quad W^\top W = I_r. \qquad (4.12) \\
C^{new} &= \underset{C}{\mathrm{argmax}} -\frac{1}{2}\mathrm{Tr}\Big(\big(\mathbf{Y}^\top\mathbf{Y} - 2\mathbf{Y}^\top\mathbb{E}\big(\mathbf{T} \mid \mathbf{X}, \mathbf{Y}, \theta^{old}\big)C^\top + C\mathbb{E}\big(\mathbf{T}^\top\mathbf{T} \mid \\
&\qquad \mathbf{X}, \mathbf{Y}, \theta^{old}\big)C^\top\big)\Psi_f^{old^{-1}}\Big) \qquad \text{s.t.} \quad C^\top C = I_r. \qquad (4.13)
\end{aligned}$$

Under the original and simpler PPLS model, variance matrices Ψ_e^{old} and Ψ_f^{old} are isotropic, and el Bouhaddani et al. (2018) could derive closed form expressions for W^{new} and C^{new}, by considering Lagrangian functions. However, closed form expressions can not be derived from the Lagrangians in our case, and so optimizations over the Stiefel Manifolds $\{W \in \mathbb{R}^{p\times r} \mid W^\top W = I_r\}$ and $\{C \in \mathbb{R}^{q\times r} \mid C^\top C = I_r\}$ have to be numerically performed (Siegel, 2019, Wen and Yin, 2010) to update W^{new} and C^{new} in each M-step of the EM algorithm, which is computationally intensive.

4.E Appendix: Additional results under the second simulation study

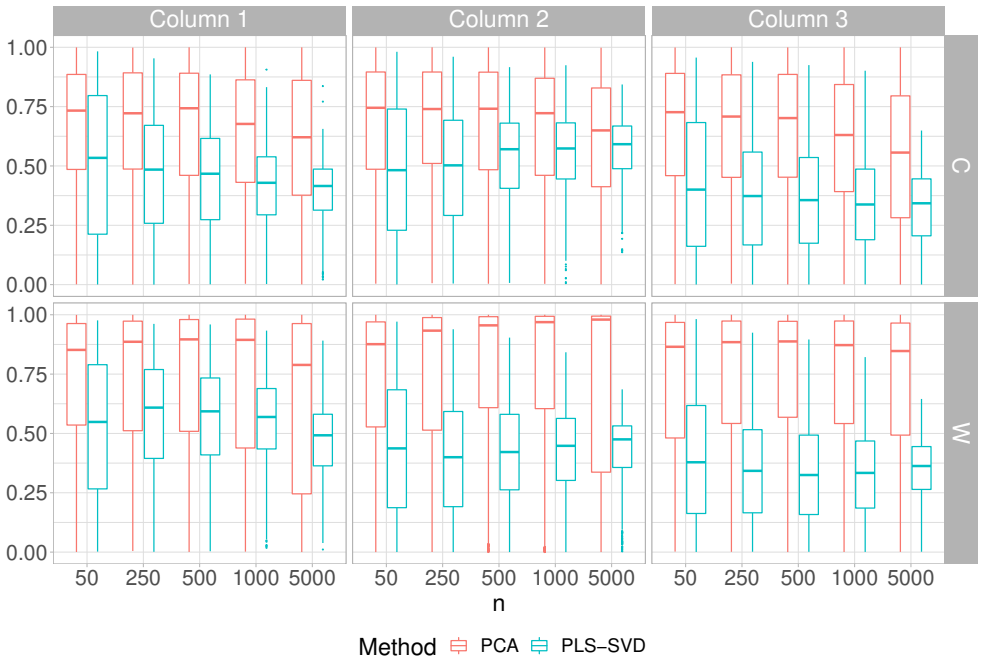


Figure 4.4: Distribution of the absolute values of the cosine similarity between the columns of the weight matrices estimated with the PPLS EM algorithm and the ones obtained via (i) two distinct PCAs on $\mathbf{X}$ and $\mathbf{Y}$, and (ii) PLS-SVD on $(\mathbf{X}, \mathbf{Y})$. The results are computed over 1000 simulations, for $p = q = 20$, $r = 3$, and different sample sizes $n \in \{50, 250, 500, 1000, 5000\}$. The top panels correspond to weights C, and the bottom panels correspond to the weights W.

Chapter 5

On the use of cross-validation for the calibration of the tuning parameter in the adaptive lasso

This Chapter corresponds to the preprint available at https://arxiv.org/abs/2005.10119, and written with Nadim Ballout and Vivian Viallon.

Abstract

The adaptive lasso refers to a class of methods, which were shown to generally enjoy better theoretical and empirical performance, at no additional computational cost, compared with the original lasso. As a result, it is very popular in practice. It relies on a weighted version of the L_1-norm penalty, where weights are typically derived from an initial estimator of the parameter vector. Irrespective of the method chosen to obtain this initial estimator, the performance of the corresponding version of the adaptive lasso critically depends on the value of the tuning parameter, which controls the magnitude of the weighted L_1-norm in the penalized criterion. In this article, we show that the standard cross-validation, although very popular in this context, has a severe defect when applied for the calibration of the tuning parameter in the adaptive lasso. We further propose a simple cross-validation scheme which corrects this defect. Empirical results from a simulation study confirms the superiority of our approach, in terms of both support recovery and prediction error, for several versions of the adaptive lasso, including the popular one-step lasso. Although we focus on the adaptive lasso under linear regression models, our work likely extends to other regression models, as well as to the adaptive versions of other penalized approaches, such as the group lasso, fused lasso, and data shared lasso.

5.1 Introduction

High dimensional data are characterized by a number p of variables larger, or at least not significantly lower, than the sample size n. They have become ubiquitous in many fields, including biology, medicine, sociology, and economy (Giraud, 2014). Their analysis raises a number of statistical challenges (Fan and Li, 2006, Hastie et al., 2009), usually summarized under the term *curse of dimensionality*. Consequently, it has attracted a lot of attention in the statistical literature over the past decades (Bühlmann and van De Geer, 2011, Donoho et al., 2000, Fan and Li, 2006, Hastie et al., 2009, 2015). In particular, a variety of approaches based on the optimization of penalized versions of the log-likelihood have been developed, to estimate the true parameter vector $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)^T \in \mathbb{R}^p$ under high-dimensional parametric regression models (Huang et al., 2008a, Tibshirani, 1996). These approaches use a penalty term, whose strength is controlled by a tuning parameter, and which is added to the loss-function so that the estimation can take advantage of some property that the true parameter vector $\boldsymbol{\beta}^*$ is expected to fulfill. For example, when $\boldsymbol{\beta}^*$ is expected to be sparse, popular approaches rely on the use of L_q penalties, $q \leq 1$. In particular, the arguably most popular approach is the lasso, which uses an L_1-norm penalty. Extensions such as the group lasso (Jacob et al., 2009), fused lasso (Tibshirani et al., 2005), generalized fused lasso (Viallon et al., 2016), data shared lasso (Ballout et al., 2020, Ballout and Viallon, 2017, Gross and Tibshirani, 2016, Ollier and Viallon, 2017), etc., rely on structured sparsity inducing norms, and can be used when some given structured sparsity is expected in $\boldsymbol{\beta}^*$.

We will here focus on another extension of the lasso, namely the adaptive lasso (Bühlmann and Meier, 2008, Bühlmann and van De Geer, 2011, Zou, 2006). This refers to a class of methods where the L_1-norm $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^p |\beta_j|$ used in the standard lasso is replaced by a weighted version $\sum_{j=1}^p w_j|\beta_j|$. The weights w_j are typically data-driven, of the form $w_j = 1/(|\tilde{\beta}_j| + \varepsilon)$, with $\varepsilon > 0$, and where $(\tilde{\beta}_j)_{1\leq j\leq p}$ are some initial estimates of the parameters $(\beta_j^*)_{1\leq j\leq p}$. Here, we will mostly focus on three popular versions of the adaptive lasso: (i) the original adaptive lasso introduced by Zou (2006), where weights are derived from Ordinary Least Squares (OLS) estimates; (ii) the version proposed by Bühlmann and Meier (2008), where weights are computed from lasso estimates (tuned in a prediction optimal way with the tuning parameter selected by cross-validation; we will get back to this particular point in more details below); and (iii) the version proposed by Zhang et al. (2008), where weights are computed from ridge estimates (again, tuned in a prediction optimal way). In the rest of this article, the original adaptive lasso introduced by Zou (2006) will be referred to as the ols-adaptive lasso, the one proposed by Zhang et al. (2008) as the ridge-adaptive lasso, while we will refer to the method described by Bühlmann and Meier (2008) as the one-step lasso, following their terminology. The popularity of the adaptive lasso can be explained as follows. First, it can be implemented very easily and efficiently using algorithms originally developed for the lasso, such as the

`glmnet` R package (Friedman et al., 2010). Second, it has been shown to usually outperform the lasso. For example, in the fixed p case, Zou (2006) established that the lasso estimates do not enjoy the asymptotic oracle property (in the sense of Fan and Li (2006)), while the ols-adaptive lasso estimates do under mild conditions on the tuning parameter. In addition, conditions ensuring support recovery in the non-asymptotic framework (which especially allows the study of the $p \gg n$ case) are weaker for the one-step lasso than for the lasso; see, e.g., Corollaries 7.8-7.9 and Section 2.8.3, of Bühlmann and van De Geer (2011).

However, as for other penalized approaches, the theoretical and empirical performance of the adaptive lasso critically depends on the value of the tuning parameter. Its theoretically optimal value involves unknown quantities, such as the variance of the noise under linear regression models, but also quantities related to the compatibility or irrepresentability condition (Bühlmann and van De Geer, 2011). Consequently, the practical selection, or calibration, of the tuning parameter also has attracted a lot of attention in the statistical literature (Arlot, 2019, Chen and Chen, 2008, Chichignoud et al., 2016, Giacobino et al., 2017). A simple and popular strategy relies on cross-validation (Allen, 1974, Hastie et al., 2009, Stone, 1974). In particular, the K-fold cross-validation (Geisser, 1975) is implemented in many publicly available lasso solvers, such as the `glmnet` R package (Friedman et al., 2010), for the calibration of the tuning parameter. Moreover, it is the method that Bühlmann and Meier (2008) recommend for the calibration of the tuning parameters in the one-step lasso, for both the initial and final estimators (see below for more details). In the present article, we will describe a defect of the standard K-fold cross-validation when used to calibrate the tuning parameter in the adaptive lasso. We will then present a simple alternative cross-validation scheme, which rectifies this flaw.

The rest of the article is organized as follows. In section 5.2, we start with a brief overview on the principles of the adaptive lasso. Then, we illustrate the flaw of the standard K-fold cross-validation when used to calibrate the tuning parameter in the adaptive lasso, and describe one simple solution to rectify this defect. In Section 5.4, we present results from a comprehensive simulation study where we empirically establish the superiority of our proposal over the standard one. Concluding remarks are given in Section 5.5.

5.2 The adaptive lasso under the linear regression model

5.2.1 Main notation and working model

As above, we will denote the sample size by n, and the number of covariates by p. For simplicity, we will focus on linear regression models of the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \xi, \tag{5.1}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ is the response vector, $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)^T \in \mathbb{R}^{n \times p}$ is the design matrix, $\boldsymbol{\beta}^* = (\beta_1^*, ..., \beta_p^*)^T \in \mathbb{R}^p$ is the p-dimensional vector of unknown parameters to be estimated, and $\xi \in \mathbb{R}^n$ is some random noise. We further denote the support of $\boldsymbol{\beta}^*$ by $J = \{j : \beta_j^* \neq 0\}$.

For any positive integer $d \geq 1$, and any vector $\mathbf{u} \in \mathbb{R}^d$, we will denote the usual Euclidian norm (or L_2-norm) by $\|\mathbf{u}\|_2$. We will let $\mathbf{0}_d$ and $\mathbf{1}_d$ be the vectors of size d with components all equal to 0 and 1 respectively, and $\mathbf{I}_d$ be the $d \times d$ identity matrix. For any real matrix $\mathbf{M} = (M_1, \ldots, M_d) \in \mathbb{R}^{n \times d}$, and any subset $E \subseteq \{1, \ldots, d\}$, $\mathbf{M}_E$ will denote the submatrix composed of the columns $(M_j)_{j \in E}$. We will further denote the cardinality of E by $|E|$.

Finally, for any sample $D_0 = \{y_i, \mathbf{x}_i\}_{i \in I_0}$, with I_0 a given set of integers, and any estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}^*$, we will denote by

$$\text{Pred.Error}(D_0, \hat{\boldsymbol{\beta}}) = \frac{1}{|I_0|} \sum_{i \in I_0} (y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}})^2,$$

the prediction error corresponding to $\hat{\boldsymbol{\beta}}$ evaluated on the sample D_0.

5.2.2 The lasso and adaptive lasso

For any $\lambda \geq 0$, the lasso estimator $\hat{\boldsymbol{\beta}}_{\text{lasso}}(\lambda)$ (Tibshirani, 1996) is defined as any minimizer over $\boldsymbol{\beta} \in \mathbb{R}^p$ of the penalized criterion

$$\frac{||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2}{n} + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{5.2}$$

The tuning parameter λ controls the amount of regularization through the L_1-norm $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^p |\beta_j|$. In practice, an appropriate value for this parameter has to be used to guarantee good statistical performance for $\hat{\boldsymbol{\beta}}_{\text{lasso}}(\lambda)$, with respect to both support recovery and prediction accuracy. As mentioned above, a popular strategy relies on K-fold cross-validation (Hastie et al., 2009) whose pseudo-code is recalled in Algorithm 3 in Appendix 5.A. Let λ^{CV} be the value of λ selected by K-fold cross-validation, and let $\hat{\boldsymbol{\beta}}_{\text{lasso}}^{\text{CV}} = \hat{\boldsymbol{\beta}}_{\text{lasso}}(\lambda^{\text{CV}})$ denote one solution of (5.2) with λ set to λ^{CV}.

Now, denote by $\mathbf{w} = (w_1, \ldots, w_p) \in \mathbb{R}_{\geq 0}^p$ any given vector of non-negative weights. For any $\lambda \geq 0$, the adaptive lasso estimator $\hat{\boldsymbol{\beta}}_{\text{ada}}(\lambda; \mathbf{w})$ is defined (Zou, 2006) as any minimizer over $\boldsymbol{\beta} \in \mathbb{R}^p$ of the criterion

$$\frac{||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2}{n} + \lambda \sum_{j=1}^{p} w_j |\beta_j|. \tag{5.3}$$

Of course, the adaptive lasso reduces to the standard lasso for the particular choice of the weight vector $\mathbf{w} = \mathbb{1}_p$: any solution $\hat{\boldsymbol{\beta}}_{\text{lasso}}(\lambda)$ is also a solution $\hat{\boldsymbol{\beta}}_{\text{ada}}(\lambda; \mathbb{1}_p)$. In practice, the

weights are usually set to $w_j = 1/|\tilde{\beta}_j|$, or $w_j = 1/(|\tilde{\beta}_j| + \varepsilon)$, with $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, ..., \tilde{\beta}_p)$ an initial estimator of $\boldsymbol{\beta}^*$, and ε some non-negative real number. A positive value for the ε parameter guarantees that every component of the weight vector is finite, so that every component of $\hat{\boldsymbol{\beta}}_{\text{ada}}(\lambda; \mathbf{w})$ has a chance to be non-zero. If the initial estimator is good enough, then $\tilde{\beta}_j$ is close to 0 for $j \notin J$, and less so for $j \in J$: if, in addition, ε is null or close enough to 0, weights w_j are large for $j \notin J$, and less so for $j \in J$. Then, components $j \notin J$ of the parameter vector are more heavily penalized than the components $j \in J$. Various initial estimates can be used to derive the weight vector $\mathbf{w}$. When $p < n$, Zou (2006) suggests the use of $\mathbf{w}_{\text{OLS}} = 1/|\tilde{\boldsymbol{\beta}}_{\text{OLS}}|$, where $\tilde{\boldsymbol{\beta}}_{\text{OLS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ is the OLS estimator. As mentioned above, we will refer to the corresponding approach as the ols-adaptive lasso. In the fixed p case, Zou (2006) established the asymptotic oracular property for the ols-adaptive lasso under mild conditions, according to which the ols-adaptive lasso is consistent in terms of variable selection (or sparsistent), and the distribution of $\left(\hat{\beta}_{\text{adap,j}}(\lambda; \mathbf{w}_{\text{OLS}})\right)_{j \in J}$ is Gaussian with the same expectation and covariance matrix than that of $(\mathbf{X}_J^T\mathbf{X}_J)^{-1}\mathbf{X}_J^T\mathbf{y}$, the OLS estimator that would been obtained if J were known in advance. The ols-adaptive lasso inherits these good properties from the $\sqrt{n}$-consistency of the OLS estimate $\tilde{\boldsymbol{\beta}}_{\text{OLS}}$ in a low-dimensional setting. However, the OLS estimate is less attractive when $p > n$. In such high-dimensional settings, Zhang et al. (2008) suggested the use of ridge regression for the computation of the weights $\mathbf{w}_{\text{ridge}} = 1/|\hat{\boldsymbol{\beta}}^{\text{CV}}_{\text{ridge}}|$. Here, $\hat{\boldsymbol{\beta}}^{\text{CV}}_{\text{ridge}}$ denotes the estimator returned by a ridge regression with tuning parameter selected by K-fold cross-validation; we recall that the ridge regression is a penalized approach similar to the lasso, but where the L_1-norm $\|\boldsymbol{\beta}\|_1$ used in the penalty is replaced by the squared L_2-norm $\|\boldsymbol{\beta}\|_2^2$ (Hoerl and Kennard, 1970). In the rest of the article, we will refer to this particular method as the ridge-adaptive lasso. Alternatively, Bühlmann and Meier (2008) suggest the use of weights $\mathbf{w}_{1-\text{step}}$ derived from $\hat{\boldsymbol{\beta}}^{\text{CV}}_{\text{lasso}}$. This leads to what they refer to as the one-step lasso. Thanks to the so-called screening property of $\hat{\boldsymbol{\beta}}^{\text{CV}}_{\text{lasso}}$, the one-step lasso has been shown to be sparsistent under weaker irrepresentability conditions than those required for the lasso (Bühlmann and van De Geer, 2011). We shall mention that other choices for the weights have been proposed in the literature: for example, the use of univariate OLS estimators was suggested by Huang et al. (2008b).

5.2.3 K-fold cross-validation for the calibration of the tuning parameter in the lasso and the adaptive lasso

Several versions of cross-validation have been proposed in the literature, but the K-fold version is arguably the most popular one in practice (Hastie et al., 2009). It first relies on partitioning the original sample $D = (y_i, \mathbf{x}_i)_{1 \leq i \leq n}$ into $K \geq 2$ balanced folds $D^{(1)}, \ldots, D^{(K)}$, with $D = \cup_{k=1}^{K} D^{(k)}$. The cross-validation then consists of K steps: at each step k, (i) the fold $D^{(k)}$ is used as an "independent" test sample, while the remaining $K - 1$ folds $D \setminus D^{(k)}$ are combined and jointly used as the training sam-

ple, (ii) the estimator $\hat{\boldsymbol{\beta}}_k$ is constructed on the training sample $D \setminus D^{(k)}$, and (iii) its prediction error $\text{Pred.Err}(D^{(k)}, \hat{\boldsymbol{\beta}}_k)$ is evaluated on the test sample $D^{(k)}$. The cross-validated prediction error is finally defined as the average of these K prediction errors: $(1/K) \times \sum_k \text{Pred.Err}(D^{(k)}, \hat{\boldsymbol{\beta}}_k)$.

This cross-validated prediction error can be used to assess the predictive performance of estimators, and to compare the predictive performance among a set of estimators. In particular, for a given weight vector $\mathbf{w} \in \mathbb{R}_{\geq 0}$, it is commonly used to compare the predictive performance of the set of estimators $(\hat{\boldsymbol{\beta}}_{\text{ada}}(\lambda_r; \mathbf{w}))_{1 \leq r \leq R}$, for any given sequence $\Lambda = (\lambda_1, \ldots, \lambda_R)$ of candidate values for the tuning parameter. It can then be used to select the optimal tuning parameter value, say, $\lambda_{\text{CV}}(\mathbf{w})$, and equivalently, the corresponding optimal estimator $\hat{\boldsymbol{\beta}}_{\text{ada}}(\lambda_{\text{CV}}(\mathbf{w}); \mathbf{w})$. See the pseudo-code given in Algorithm 3 in Appendix 5.A for the detailed description of the cross-validation in this particular setting.

The cross-validated prediction error is known to have some limitations; see, e.g., Chapter 7 of Hastie et al. (2009). See also Arlot (2008), Arlot and Celisse (2010). However, it is usually considered to perform reasonably well in practice, and is therefore still very popular, in particular for the calibration of the tuning parameter of the lasso and the adaptive lasso. As mentioned in the Introduction 5.1, it is for instance available in packages like `glmnet` (Friedman et al., 2010), and it is also the method that Bühlmann and Meier (2008) used for the selection of the tuning parameter for both the initial and final estimators in their one-step lasso procedure. However, we observed a severe defect of K-fold cross-validation in the case of the adaptive lasso, which, to the best of our knowledge, has been ignored in the literature so far. Below, we present results from a simple simulation study, whose main objective is to illustrate (i) the good performance of the cross-validation when used for the calibration of the tuning parameter in the lasso, and (ii) its poor performance when applied for the calibration of the tuning parameter in the adaptive lasso. Results from a more comprehensive simulation study will be presented in Section 5.4.

In this first simple synthetic example, we generate one sample $D = (y_i, \mathbf{x}_i)_{1 \leq i \leq n}$, made of $n = 1,000$ observations under the linear regression model given in Equation (5.1) with $p = 1,000$. We first set $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)$ with $\beta_j^* = 0$ for all $j \geq 11$, and $\beta_j^* = \iota_j 0.5$ for all $j \leq 10$, where ι_j is a $\{-1, 1\}$-binary random variable, with $\mathbb{P}(\iota_j = 1) = 1/2$. Then, for each $i = 1, \ldots, n$, we generate a Gaussian random noise $\xi_i \sim N(0, 1)$, a Gaussian vector of covariates $\mathbf{x}_i = (x_{i,1}, ..., x_{i,p}) \sim N(\mathbf{0}_p, \mathbf{I}_p)$, and finally the outcome $y_i = \mathbf{x}_i^T \boldsymbol{\beta}^* + \xi_i$. Similarly, we generate one independent test sample $\mathscr{D} = (y_i, \mathbf{x}_i)_{n+1 \leq i \leq n+N}$, made of $N = 10,000$ observations drawn under the same linear model. Then, for any particular weight vector $\mathbf{w}$, the `glmnet` R package is used to compute the (adaptive) lasso estimator on an appropriate sequence $\Lambda = \Lambda(D, \mathbf{w}) = (\lambda_1(D, \mathbf{w}), \ldots, \lambda_{100}(D, \mathbf{w}))$ of 100 decreasing values for the tuning parameter (these values are equally spaced on a log-scale, and are

automatically selected, in a data-specific way, by the `glmnet` and `cv.glmnet` functions of the `glmnet` package). The `glmnet` R package is used to compute the 10-fold cross-validated prediction error for each of the 100 corresponding estimators as well. Lastly, the "true" prediction error of each estimator is approximated by its prediction error evaluated on the independent test set $\mathscr{D}$.

Given the relatively high-dimensional setting of this first simulation study, the ols-adaptive lasso is not considered here. Figure 5.1 presents our results for the lasso ($\mathbf{w} = \mathbb{1}_p$), the one-step lasso ($\mathbf{w} = \mathbf{w}_{1-\text{step}} = 1/(|\hat{\boldsymbol{\beta}}^{\text{CV}}_{\text{lasso}}| + 10^{-4})$), and the ridge-adaptive lasso ($\mathbf{w} = 1/(|\hat{\boldsymbol{\beta}}^{\text{CV}}_{\text{ridge}}| + 10^{-4})$). For the latter two, we also estimate the cross-validated prediction error using our nested cross-validation scheme, which we will introduce in the next Section. For comparability, the x-axis corresponds to the tuning parameter sequence represented as a fraction of the data-specific maximal value $\lambda_1(D, \mathbf{w})$.

First consider the standard lasso (left panel). In this case, the cross-validated prediction error does a fairly good job in approximating the true prediction error on a wide range of λ-values. In particular, the λ-value at which the cross-validated prediction error is minimized (vertical dotted red line) is very close to that at which the true prediction error is minimized (vertical dotted blue line). Moreover, the true prediction error evaluated at these two λ-values (horizontal dotted red and blue lines, respectively) are indistinguishable on this plot: in this example, the lasso estimator $\hat{\boldsymbol{\beta}}^{\text{CV}}_{\text{lasso}}$ selected by cross-validation is therefore nearly optimal with respect to prediction error. However, the cross-validated prediction error does not perform that well for the two versions of the adaptive lasso presented in this example. In particular, for the one-step lasso, the cross-validated prediction error constantly decreases as the tuning-parameter decreases, a behavior that we observed on numerous other simulation designs as well (results not shown). Then, the λ-value at which the cross-validation prediction error is minimized (vertical dotted red line) is very different from the one that minimizes the true prediction error. This suggests that the support of the one-step lasso estimator might be too large if the tuning parameter is selected via the standard cross-validation (as will be confirmed in the more comprehensive simulation study presented in Section 5.4). Moreover, the true prediction error (estimated on the test sample $\mathscr{D}$) evaluated at these two λ-values (horizontal dotted red and blue lines, respectively) are quite different: the one-step-lasso estimator selected by cross-validation is far from optimal with respect to prediction error on this example. A similar, although less pronounced, behavior is observed in the case of the ridge-adaptive lasso, suggesting that standard K-fold cross-validation is not recommended for the calibration of the tuning parameter of the adaptive lasso. On the other hand, using our proposed nested scheme, which we will introduce in the next Section, seems to correct this defect. For both the one-step lasso and ridge-adaptive lasso, the λ-value at which the corresponding cross-validated prediction error is minimized (vertical dotted green line) is close to the one minimizing the true prediction error, and the true prediction error

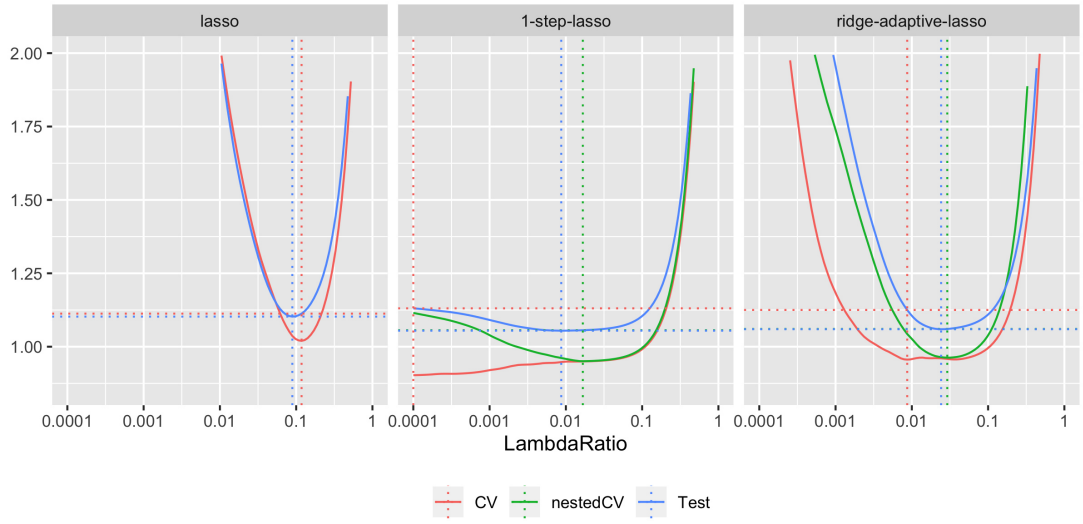


Figure 5.1: Comparison between the "true" prediction error (estimated on a truly independent test set, in blue) and the cross-validated prediction error (in red), for the lasso (left), the one-step-lasso (middle), and the ridge-adaptive lasso (right). For the latter two, the cross-validated prediction error estimated using our nested cross-validation scheme is also presented (in green). Vertical dotted lines represent the value of the tuning parameter for which each particular curve is minimized (red: standard cross-validated prediction error; green: cross-validated prediction error using our proposed nested scheme; blue: "true" prediction error). Horizontal dotted lines represent the value of the "true" prediction error for these particular values of the tuning parameter.

evaluated at this λ-value (horizontal dotted green line) is close to the optimal prediction error: the one-step-lasso and the ridge-adaptive lasso estimators selected by our proposed nested cross-validation scheme are both nearly optimal with respect to prediction error on this example.

5.3 A new cross-validation scheme for the adaptive lasso

5.3.1 Additional notation

For any sample of observations $D_0 = \{y_i, \mathbf{x}_i\}_{i \in I_0}$, with I_0 a given set of integers, any vector of non-negative weights $\mathbf{w}$, and any non-negative λ value, we let $\text{Lasso}(D_0, \mathbf{w}, \lambda)$ denote one particular solution of the adaptive lasso (5.3), when computed on sample D_0 with weights $\mathbf{w}$ and tuning parameter λ. For any positive integer $K \geq 2$, let $\text{CVLasso}(D_0, \mathbf{w}, K)$ be an adaptive lasso estimator computed on D_0 with weights $\mathbf{w}$, and with a tuning parameter set to its optimal value according to the standard K-fold cross-validation. Similarly, $\text{nestedCVLasso}(D_0, \mathbf{w}, K)$ will denote an adaptive lasso estimator computed on D_0 with weights $\mathbf{w}$, but this time with a tuning parameter set to its optimal value according to our proposed K-fold cross-validation scheme, which will be introduced below. In the the presentation of the ridge-adaptive lasso, we will further use

the shorthand CVRidge(D_0, K) to denote the ridge estimator (that will be used to compute the weights in the ridge-adaptive lasso) computed on D_0 with a tuning parameter set to its optimal value according to the standard K-fold cross-validation.

5.3.2 Our proposal

We start by giving some intuition about the defect of the standard K-fold cross-validation when used for the calibration of the tuning parameter in the adaptive lasso, as described in Section 5.2.3. Recall that the overall principle of cross-validation is to mimic "independent" test samples. Then, for the cross-validation to perform well, at each step k of the K-fold cross-validation, the whole estimation procedure should be performed on the training sample $D \setminus D^{(k)}$, and should not use any information from the test sample $D^{(k)}$. However, the weights used in the adaptive lasso are derived from initial estimates computed on the entire original sample D. Therefore, considering the estimation of the adaptive lasso estimator as a whole, it does use information from the test samples: the independence of these test samples is not guaranteed, which may be the cause of the poor performance of the cross-validation in this particular framework.

Our overall proposal for the calibration of the tuning parameter, and eventually the selection of the optimal adaptive lasso estimator, is described in Algorithm 1. The only difference with the usual approach lies in Step 2. Usually, this step consists of Step 2-a, where the standard cross-validation CVLasso$(D, \mathbf{w}, K)$ is used. Our proposal consists in replacing it by Step 2-a', where our proposed nestedCVLasso$(D, \mathbf{w}, K)$, which is detailed in Algorithm 2 in the particular case of the one-step lasso, is used instead. The key difference between CVLasso (see Algorithm 3 in Appendix 5.A) and nestedCVLasso is highlighted in blue in Algorithm 2: at each step k of our proposed K-fold cross-validation, the weights used for the adaptive lasso are first recomputed on the "training" sample $D \setminus D^{(k)}$, so that the whole estimation of the adaptive lasso estimator uses information from the training sample only, before computing the corresponding prediction error on the "independent test" sample $D^{(k)}$. In the case of the one-step lasso (or the ridge-adaptive lasso), this leads to a nested cross-validation scheme. More precisely, an appropriate sequence of candidate λ values for the tuning parameter has first to be chosen, which typically depends on the weights computed on the whole original sample (we briefly get back to this point below). Then, at each step k of the "outer" K-fold cross-validation, one "inner" standard cross-validation is performed to compute the optimal lasso (or ridge) estimator on the training sample $D \setminus D^{(k)}$, from which the weights are derived, before the corresponding adaptive lasso estimator is computed for each of the λ values of the sequence, and their predictive performance is eventually evaluated on the test sample $D^{(k)}$. For each λ value of the sequence, the predictive performance is then averaged over the K folds. The optimal value for the tuning parameter is defined as the value that minimizes this averaged criterion. The optimal adaptive lasso estimator finally corresponds to the

adaptive lasso computed on D, with weights $\mathbf{w}$ also computed on D, and tuning parameter set to this optimal value.

Algorithm 1: Cross-validation for the adaptive lasso. Version (i) of step (1-a) corresponds to the one-step lasso, while version (ii) corresponds to the ols-adaptive Lasso, and version (iii) to the ridge-adaptive lasso. The usual approach corresponds to the algorithm ran with Step (2-a), while we propose to use Step (2-a') instead, which is further detailed in Algorithm 2.

Data: Sample: $D = \{y_i, \mathbf{x}_i\}_{i=1}^n$, Version of the adaptive lasso, Number of folds: K, and parameter $\varepsilon \geq 0$

Result: $\hat{\boldsymbol{\beta}}^{\text{CV}}_{\text{ada}}$

Step 1: Computation of the initial estimates and weights;

(1-a) *Initial estimates: either (i), (ii) or (iii) below, depending on the considered version of the adaptive lasso* ;

(i): $\tilde{\boldsymbol{\beta}} = \text{CVLasso}(D, \mathbb{1}_p, K)$ `/* one-step lasso */`

(ii): $\tilde{\boldsymbol{\beta}} = \text{OLS}(D)$ `/* ols-adaptive lasso */`

(iii): $\tilde{\boldsymbol{\beta}} = \text{CVRidge}(D, K)$ `/* ridge-adaptive Lasso */`

(1-b) *Weights*;

$\mathbf{w} = 1/(|\tilde{\boldsymbol{\beta}}| + \varepsilon)$;

Step 2: Computation of the final estimates;

(2-a) $\hat{\boldsymbol{\beta}}^{\text{CV}}_{\text{ada}}(\mathbf{w}) = \text{CVLasso}(D, \mathbf{w}, K)$;

(2-a') $\hat{\boldsymbol{\beta}}^{\text{CV}}_{\text{ada}}(\mathbf{w}) = \text{nestedCVLasso}(D, \mathbf{w}, K)$;

A first remark is that our proposal involves, as a preliminary step, the computation of an appropriate sequence of candidate tuning parameters on the entire original sample, and based on the weights $\mathbf{w}$ that are also computed on the entire original sample. In other words, our claim above that, with our proposal, the whole estimation process is independent from the test samples was actually overstated. However, it is not clear how to correct this slight violation of the independence of the test samples. Moreover, results from our simulation study suggest that this violation seems to be of no practical consequence.

We shall further stress that when applied in the case of the ols-adaptive lasso, our proposal cannot be seen as a nested cross-validation anymore. Indeed, in this case, there is no inner cross-validation needed to compute the optimal OLS estimator on the training sample at each step of the cross-validation (since the OLS does not rely on any hyper-parameter to be optimized).

Algorithm 2: Our proposed K-fold cross-validation scheme (nestedCVLasso) for the calibration of the tuning parameter of the adaptive lasso: the case of the one-step lasso.

Data: Sample: $D = \{y_i, \mathbf{x}_i\}_{i=1}^n$, Weights: $\mathbf{w}$, Number of folds: K, and parameter $\varepsilon \geq 0$

Result: $\hat{\boldsymbol{\beta}}_{\text{ada}}^{nCV}(\mathbf{w}) = \hat{\boldsymbol{\beta}}_{\text{ada}}(\lambda^{nCV}, \mathbf{w}, D)$

Computation of a sequence $\Lambda := \Lambda(D, \mathbf{w}) = (\lambda_1, \ldots, \lambda_R)$;

Division of D into K folds: $D = \cup_{k=1}^K D^{(k)}$;

for $k \in \{1, \ldots, K\}$ **do**

  /* Computation of the weights on $D \setminus D^{(k)}$ */

  $\tilde{\boldsymbol{\beta}}^{(k)} = \text{CVLasso}(D \setminus D^{(k)}, \mathbb{1}_p, K)$;

  $\mathbf{w}_k = 1/(|\tilde{\boldsymbol{\beta}}^{(k)}| + \varepsilon)$;

  for $r \in \{1, \ldots, R\}$ **do**

    $\hat{\boldsymbol{\beta}}_{r,k} = \text{Lasso}(D \setminus D^{(k)}, \mathbf{w}_k, \lambda_r)$;

    $E_{r,k} = \text{Pred.Error}(D^{(k)}, \hat{\boldsymbol{\beta}}_{r,k})$;

  end

end

$r^* = \text{argmin}_r \left\{ \sum_{k=1}^K E_{r,k} \right\}$;

$\lambda^{nCV} = \lambda_{r^*}$;

$\hat{\boldsymbol{\beta}}_{\text{ada}}^{nCV}(\mathbf{w}) = \text{Lasso}(D, \mathbf{w}, \lambda^{nCV})$;

5.4 Simulation study

We now present results from a more comprehensive simulation study, which extends the simple one presented in Section 5.2.3. As before, we generate a sample $D = (y_i, \mathbf{x}_i)_{1 \leq i \leq n}$, made of $n = 1,000$ observations drawn under the linear regression model (5.1). Here, we make the number of covariates p vary in $\{100, 500, 1000\}$, and the number of relevant covariates p_0 vary in $\{10, 50\}$. We randomly select the support J of $\boldsymbol{\beta}^*$, with $|J| = p_0$. Then, we set $\beta_j^* = 0$ for all $j \notin J$, and $\beta_j^* = \iota_j \beta$, for all $j \in J$, where ι_j is a $\{-1, 1\}$-binary variable, with $\mathbb{P}(\iota_j = 1) = 1/2$. As for the signal strength β, we make it vary in $\{1/4, 1/2, 1, 3/2\}$. Then, for each $i = 1, \ldots, n$, we generate a Gaussian random noise $\xi_i \sim N(0, 1)$, a Gaussian vector of covariates $\mathbf{x}_i = (x_{i,1}, ..., x_{i,p}) \sim N(\mathbf{0}_p, \mathbf{I}_p)$, and finally the outcome $y_i = \mathbf{x}_i^T \boldsymbol{\beta}^* + \xi_i$. Additionally, we generate one independent test sample $\mathscr{D} = (y_i, \mathbf{x}_i)_{n+1 \leq i \leq n+N}$, made of $N = 10,000$ observations drawn under the same linear model.

We consider the one-step lasso and the ridge-adaptive lasso as before. The ols-adaptive

lasso is also considered in the low-dimensional scenario where $p = 100$. As in our previous simple example, weights were derived from initial estimates $\tilde{\boldsymbol{\beta}}$ as $w_j = 1/(|\tilde{\beta}_j| + 10^{-4})$; other choices for the parameter $\varepsilon \in \{0, 10^{-6}, 10^{-2}\}$ were tested, and led to very similar results (not shown). For each method, the optimal adaptive lasso estimator is computed following either the standard 10-fold cross-validation (Algorithm 1 with Step 2-a: we will refer to these estimators as the one-step lasso CV, the ridge-adaptive lasso CV, and the ols-adaptive lasso CV respectively) or our proposed 10-fold "nested" cross-validation scheme (Algorithm 1 with Step 2-a': we will refer to these estimators as the one-step lasso nested CV, the ridge-adaptive lasso nested CV, and the ols-adaptive lasso nested CV, respectively). For comparison, results from the lasso estimator, with tuning parameter selected via standard 10-fold cross-validation (see Algorithm 3 in Appendix 5.A) are presented; we will refer to this estimator as the lasso CV. Functions from the `glmnet` R package are used to compute these different estimators. We evaluate both the (signed) support accuracy and the prediction error attached to each estimator $\hat{\boldsymbol{\beta}}$. More precisely, the signed support accuracy is defined as $\{\sum_{j=1}^{p} \mathbb{1}(\text{sign}(\beta_j^*) = \text{sign}(\hat{\beta}_j))\}/p$, where $\mathbb{1}$ is the indicator function, and sign the sign function, that is $\text{sign}(x) = +1$ if $x > 0$, $\text{sign}(x) = -1$ if $x < 0$, and $\text{sign}(x) = 0$ if $x = 0$. As for the prediction error, we simply compute $\text{Pred.Err}(\mathscr{D}, \hat{\boldsymbol{\beta}})$, as before. These two criteria are averaged over the 50 replications we consider for each combination of values for the parameters (p, p_0, β).

Figure 5.2 presents the results. First, they illustrate that the adaptive lasso usually outperforms the lasso, in terms of prediction error and support accuracy. Second, focusing on the adaptive lasso estimators, they also illustrate that our proposal yields better performance than the standard K-fold cross-validation, in terms of both prediction error and support accuracy. This increased performance is particularly substantial for low signal strength, and high dimension (number of covariates p and/or cardinality p_0 of the support J of $\boldsymbol{\beta}^*$): in such situations, it is noteworthy that, for instance, the one-step lasso CV is typically outperformed by the simple lasso CV in terms of prediction error, and does not outperform it much in terms of support accuracy, while the one-step lasso nested CV exhibits substantially higher support accuracy and lower prediction error.

5.5 Discussion-Conclusion

In this article, we described a defect of the standard K-fold cross-validation when applied for the calibration of the tuning parameter in the adaptive lasso, with emphasis on the ols-adaptive and ridge-adaptive lasso, as well as the one-step lasso. We further proposed a simple alternative which corrects this defects.

Although we focused on the K-fold cross-validation, other cross-validation schemes (Arlot and Celisse, 2010) likely suffer from a similar defect, in which case our proposal could easily be extended to these other cross-validation schemes. In addition, we here

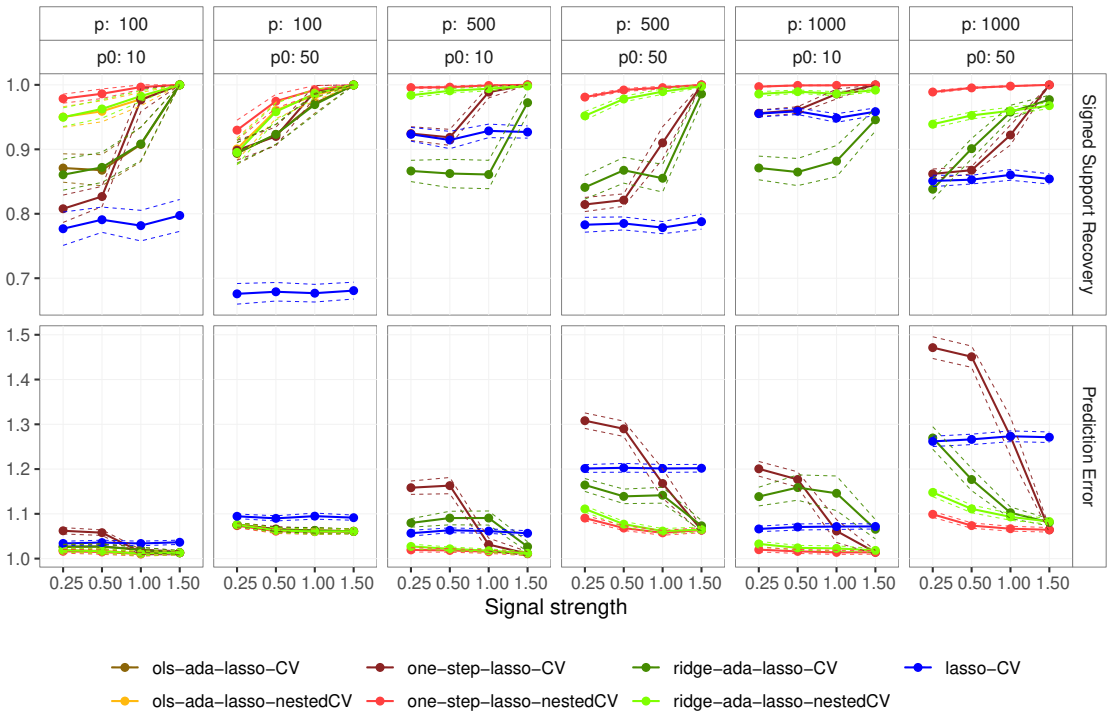


Figure 5.2: Results of the simulation study. Solid lines represent the averaged criteria, while dashed lines represent the associated 95% confidence intervals.

considered the adaptive lasso under linear regression models for simplicity, but we expect similar defects for the standard cross-validation, as well as improvements when applying appropriate extension of our proposal, for extensions of the adaptive lasso (e.g., the adaptive generalized fused lasso (Viallon et al., 2016), the adaptive data shared lasso (Ballout et al., 2020, Gross and Tibshirani, 2016, Ollier and Viallon, 2017)), and under other regression models (e.g., generalized linear models, Cox proportional hazard models).

Disclaimer

Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer /World Health Organization.

5.A Appendix: Pseudo-code of the standard K-fold cross-validation for the calibration of the tuning parameter in the (adaptive) lasso

The pseudo-code detailed in Algorithm 3 below describes the function CVLasso: when applied to the sample $D = \{y_i, \mathbf{x}_i\}_{i=1}^n$, with weights $\mathbf{w}$ and number of folds K, it produces the CV-optimal adaptive Lasso estimator $\hat{\boldsymbol{\beta}}_{\text{ada}}^{\text{CV}}(\mathbf{w}) = \hat{\boldsymbol{\beta}}_{\text{ada}}(\lambda^{\text{CV}}, \mathbf{w})$, with tuning parameter λ^{CV} set to its value minimizing the K-fold cross-validated prediction error.

Algorithm 3: The standard K-fold cross-validation CVLasso$(D, \mathbf{w}, K)$ for the calibration of the tuning parameter in the lasso and adaptive lasso

Data: Sample: $D = \{y_i, \mathbf{x}_i\}_{i=1}^n$, Weights: $\mathbf{w}$, Number of folds: K
Result: $\hat{\boldsymbol{\beta}}_{\text{ada}}^{\text{CV}}(\mathbf{w})$
Computation of a sequence $\Lambda := \Lambda(D, \mathbf{w}) = (\lambda_1, \ldots, \lambda_R)$;
Division of D into K folds: $D = \cup_{k=1}^K \{D^{(k)}\}$;
for $k \in \{1, \ldots, K\}$ **do**
    for $r \in \{1, \ldots, R\}$ **do**
        $\hat{\boldsymbol{\beta}}_{r,k} = \text{Lasso}(D \setminus D^{(k)}, \mathbf{w}, \lambda_r)$;
        $E_{r,k} = \text{Pred.Error}(D^{(k)}, \hat{\boldsymbol{\beta}}_{r,k})$;
    end
end
$r^* = \text{argmin}_r \left\{ \sum_{k=1}^K E_{r,k} \right\}$;
$\lambda^{\text{CV}} = \lambda_{r^*}$;
$\hat{\boldsymbol{\beta}}_{\text{ada}}^{\text{CV}}(\mathbf{w}) = \hat{\boldsymbol{\beta}}_{\text{ada}}(\lambda^{\text{CV}}, \mathbf{w})$;

Chapter 6

Discussion and perspectives

Cancer epidemiology is notably engaged in the study of potential causes of cancer based on observational data, including, more recently, biological mechanisms possibly involved in cancer development. In Appendix A, we present some of the tools introduced in causal inference to offer the formal framework to address these causal queries, and which was lacking with conventional tools in Statistics. In particular, counterfactual variables allow the precise definitions of causal quantities of interest, whether it is the total causal effect of an exposure, or its decomposition through natural direct and indirect effects that can be useful for the study of the biological mechanisms possibly underlying some exposure-cancer relationship. Sets of sufficient conditions have further been proposed in the literature to determine if, and if so, how, these quantities may be estimated from observational data. Yet, the practical application of causal inference and mediation analyses in cancer epidemiology faces several challenges and issues; the objective of this thesis was to explore some of them.

First, concerns have been raised in the literature regarding the relevance of causal effects estimated from observational studies for certain exposures. For example, there is no possible "direct" intervention to reduce obesity, and obesity can basically only be reduced by intervening on some of its causes, like diet or physical activity. As several practical interventions on the causes of obesity could lead to the same obesity level, this situation falls under the general case of a treatment with multiple versions. In Chapter 2, which can be seen as a complement of Petersen (2011)'s work, we focused on the particular case where the versions precede the treatment, as this situation had not been considered in depth in the literature. For the purpose of illustration, we considered the case where X, the exposure of interest, stood for the obesity at the age of 20, for which the hypothetical intervention $do(X = x)$ cannot be directly implemented in practice. Then, we investigated how the effect of this hypothetical intervention $do(X = x)$, which can still be estimated using data from a cohort study (under some assumptions), relates to the effects of interventions on its causes. In particular, we considered the situation where

some causes of X are modifiable, while others are not. As emphasized by Petersen (2011), our results state that the structure of the causal model indicates how an hypothetical intervention on X can be interpreted. Specifically here, two distinct situations may arise, depending on whether the "versions of the treatment" are relevant or not. On the one hand, focusing on the modifiable causes of X that affect Y through X only, if any, the different versions are irrelevant, and then the effect of an hypothetical intervention on X captures the effect of interventions on these causes. On the other hand, for modifiable causes W of X that affect Y not only through X (for example, if W is a confounder in the $X - Y$ relationship), the different versions are relevant, and the effect of an hypothetical intervention on X only partly captures the effect of interventions on W. In particular, we showed that under a linear model with no interaction, the effect of an hypothetical intervention on X can be seen as an indirect effect of the intervention on such causes W. Then, an interesting lead to gain further intuition and insight on the relationship between these effects would be to study other simple parametric causal models. Getting back to our motivating example where the exposure X is obesity, because most of its modifiable causes can generally be regarded as confounders in its relationship with cancer, the effect of obesity estimated from observational data is likely to differ from the effect of interventions on its causes. Simulation studies could help to gain more insight on this illustrative example.

Then, even if most causal models of interest in cancer epidemiology involve time-varying exposures, confounders and mediators, these variables are usually measured once only in practice, typically at the inclusion in the study. Practitioners then tend to overlook their time-varying nature and to work under over-simplified causal models. In Chapter 3, we investigated whether total causal effects derived when working under these simplified and generally misspecified models could still be related to longitudinal causal effects of potential interest. In particular, we focused on two situations regarding the type of available data for the exposure, and possibly mediators and confounders: when they correspond to (i) "instantaneous" levels measured at inclusion in the study or (ii) summary measures of their levels up to inclusion in the study. On the one hand, we derived sufficient conditions ensuring that the quantities estimated in practice under over-simplified causal models can be expressed as true longitudinal causal effects of interest, or some weighted averages thereof. But on the other hand, we emphasized that biases generally arise in both cases (i) and (ii) as the sufficient conditions are very restrictive, and also because the interpretability of these weighted averages is not always straightforward. It is noteworthy that we focused on the "ideal" setting where the available variables (either the variables at inclusion in the study or the summary variables) are perfectly measured, without measurement error. As our results are already mostly negative in this case, they would be even more so if the observed variables corresponded to a noisy version of the true ones. Nevertheless, in the case where the quantity estimated in practice in either situation

(i) or (ii) has a clear interpretation in terms of longitudinal causal effects of interest, such measurement error could lead to biased estimates, and would have to be taken into account. In addition, our results only cover total causal effects, while similar issues arise when considering natural direct and indirect effects, as illustrated through a few simple examples in Appendix B. Establishing a more general theory for mediation analysis, as the one presented in Chapter 3 for total causal effects, constitutes an interesting lead for future research. However, we expect the conditions to be more restrictive than for total causal effects. Overall, our results are in line with the conclusions of previous publications (Daniel et al., 2012, Maxwell and Cole, 2007, Maxwell et al., 2011), and support the need for repeated measurements for the exposure of interest, the confounders and mediators. But, because such repeated measurements are not always available, sensitivity analyses could be useful to assess the extent to which the results obtained under a given over-simplified model are biased (Ding and VanderWeele, 2016, VanderWeele and Arah, 2011). Finally, we considered the standard discrete-time framework, while a continuous-time framework might be better suited for most applications. Then, the study of the biases induced when working under discrete-time models while the true causal model is a continuous-time one, would be of interest, and several leads for future work may be considered in this regard. For example, simulation studies could be performed to illustrate the magnitude of the bias, depending on the length of the discrete time intervals and other parameters. In addition, assuming a certain "regularity" of the exposure process, mediator and confounder processes, and/or eventually a parametric model for the counterfactual outcome, we could investigate which time-lag between the measurements should be considered to prevent or at least limit bias.

We also initiated a project on high-dimensional mediation analysis, motivated by the analysis of metabolomics data for the investigation of biological mechanisms possibly involved in the obesity-cancer and other lifestyle-cancer relationships. Several approaches have been proposed in the literature for this purpose, including the latent variable model for high-dimensional mediation analysis proposed by Derkach et al. (2019). In particular, their model allows issues related to both the dimensionality of the metabolites and the large correlations among the metabolites to be tackled, by targeting a small number of uncorrelated metabolic signatures to summarize the information contained in the whole set of original metabolites. The latent variable model proposed by Derkach et al. (2019) was devised for mediation analysis with a univariate exposure (e.g. BMI), and they further used adaptive lasso penalties to encourage sparsity in the weight vector used for the construction of the metabolic signatures. Yet, as several indicators can be used to define obesity or lifestyle, we decided to extend their model to a setting where the exposure is multivariate. We decided to use different constraints for the model parameters compared to those used by Derkach et al. (2019); our choice, inspired by the constraints used by el Bouhaddani et al. (2018) for their PPLS model, notably guaranteed the identifiabil-

ity of the weight vectors up to sign. However, when studying the identifiability of our latent variable model for high-dimensional mediation analysis, we noticed a severe limitation: our model defined a subset of very particular distributions for the three sets of observed variables, under which the weight vectors to be used for the construction of the lifestyle and metabolic signatures could be obtained from much simpler models, namely two distinct PPCA models run separately on the exposure and metabolite sets. We further identified that this limitation came specifically from the constraints on the model parameters, which were too strong. We actually noticed that several other latent variable models proposed in the literature, notably the PPLS model proposed by el Bouhaddani et al. (2018), suffer from similar defects. Contrary to our model, the PPLS model proposed by el Bouhaddani et al. (2018) only involves two sets of observed variable, and we decided to focus on their simpler setting to precisely describe this limitation. In Chapter 4 we showed that the PPLS model proposed by el Bouhaddani et al. (2018) defines a very particular subset of distributions, under which the signatures with maximal covariance are necessarily of maximal variances as well. We further illustrated this limitation through simulated examples, and proposed a simple extension of their model, which is no longer restricted to particular distributions such as the ones mentioned above. Overall, our results stress that caution is needed when developing latent variable models for dimension-reduction, as they may turn out to be too simplistic when imposing too strong of constraints on the model parameters. In particular, a complex model whose structural equations seem to correctly describe the complex relationships among the observed variables, may actually define very particular distributions, under which the parameters of interest in the complex model actually reduce to some parameters in much simpler models. Interestingly, Zheng et al. (2016) already emphasized that the Probabilistic PLS Regression model proposed by Li et al. (2015) shared similarities with the probabilistic formulation of Principal Component Regression proposed by Ge et al. (2011). We plan to study in more detail the latent variable model proposed by Li et al. (2015), as well as the model proposed by Derkach et al. (2019), to precisely describe the limitation under these two models too.

Another extension of our work would consist in developing a latent variable model that is well suited for the mediation analysis framework. However, this extension is not straightforward, as it is challenging to make sure that the parameters of interest are identifiable, while ensuring they cannot be identified from only one or two of the three sets of observed variables. For example, consider the model presented in Figure 6.1. Although the latent variable $\mathbf{T}$ appears in the three structural equations that define $\mathbf{X}$, $\mathbf{M}$ and Y, it is easy to show that the first two equations are sufficient to identify the parameters W and C, which contain the weights to be used for the construction of the lifestyle and metabolic signatures. Again, at first glance, and just as the model proposed by Derkach et al. (2019), this model seems to be well tailored for the mediation analysis

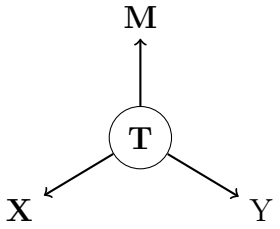
Latent variable model for mediation analysis inspired by our extension of the PPLS model proposed by el Bouhaddani et al. (2018):

$$\mathbf{X} = \mathbf{T}W^\top + \varepsilon_{\mathbf{X}},$$
$$\mathbf{M} = \mathbf{T}C^\top + \varepsilon_{\mathbf{M}}\ , \qquad Y = \mathbf{T}\gamma + \varepsilon_Y.$$

Under the assumptions:

(1) $\mathbf{T} \sim \mathcal{N}(\mathbf{0}_r, \Sigma_{\mathbf{T}})$.
(2) $\varepsilon_{\mathbf{X}} \sim \mathcal{N}(\mathbf{0}_{p_X}; \Psi_X)$. (3) $\varepsilon_{\mathbf{M}} \sim \mathcal{N}(\mathbf{0}_{p_M}; \Psi_M)$.
(4) Ψ_X is a $p_X \times p_X$ semi-positive definite matrix.
(5) Ψ_M is a $p_M \times p_M$ semi-positive definite matrix.
(6) $\varepsilon_Y \sim \mathcal{N}\left(0, \sigma_Y^2\right)$.
(7) W and C are respectively $p_X \times r$ and $p_M \times r$ semi-orthogonal matrices .
(8) $r < \min(p_X, p_M)$.

(a)



(b)

Figure 6.1: (a) A latent variable model for mediation analysis, inspired by our extension of the PPLS model proposed by el Bouhaddani et al. (2018). (b) Graphical representation of the latent variable model for mediation analysis presented in Figure 6.1 (a). Circled nodes represent sets of latent variables.

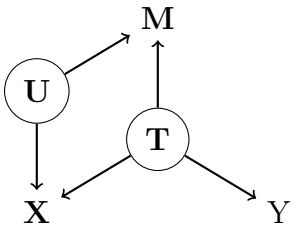
setting, but it actually defines a very particular subset of distributions for $\mathbf{X}, \mathbf{M}$ and Y, where the signatures for the $\mathbf{X}$ and $\mathbf{M}$ sets which summarize the relationships between $\mathbf{X}$, $\mathbf{M}$, and Y, are necessarily those of our proposed extension of the PPLS model run on $(\mathbf{X}, \mathbf{M})$ only. Then, more complex latent variable models are needed. In particular we can consider models with additional sets of latent variables, which would be related only to $\mathbf{X}$ and/or $\mathbf{M}$, as depicted in Figure 6.2 (b). Such models can be seen as extensions of the probabilistic PLS Regression model proposed by Zheng et al. (2016). Under a model such as the one presented in Figure 6.2 (a), parameters W and C can generally not be identified from the first two structural equations: then, the metabolic and lifestyle signatures would really depend on the full distribution of $(\mathbf{X}, \mathbf{M}, Y)$; in particular, the marginal distribution of $(\mathbf{X}, \mathbf{M})$ is not sufficient to identify W and C, and then to define the signatures. However, establishing the identifiability of the model parameters is not straightforward under these models. Moreover, the implementation of the (penalized) EM algorithm to estimate the model parameters could be challenging too. Indeed, we noted that the implementation of the EM algorithm for our extension of the PPLS model was already less straightforward than for the initial model proposed by el Bouhaddani et al. (2018). In particular, this algorithm required additional numerical optimization steps for which we have not found any efficient method yet. Then the estimation procedure would likely be even more complicated for a latent variable model for mediation analysis such as the one given in Figure 6.2 (a). Alternatively, a different direction could be considered. For example, the approach proposed by Geuter et al. (2020), which directly looks for

Latent variable model for mediation analysis inspired by our extension of the PPLS model proposed by el Bouhaddani et al. (2018):

$$\mathbf{X} = \mathbf{T}W^\top + \mathbf{U}A^\top + \varepsilon_{\mathbf{X}},$$
$$\mathbf{M} = \mathbf{T}C^\top + \mathbf{U}B^\top + \varepsilon_{\mathbf{M}}\ , \qquad Y = \mathbf{T}\gamma + \varepsilon_Y.$$

Under the assumptions:

(1) $\mathbf{T} \sim \mathcal{N}(\mathbf{0}_r, \Sigma_{\mathbf{U}})$. (2) $\mathbf{U} \sim \mathcal{N}(\mathbf{0}_s, \Sigma_{\mathbf{T}})$.
(3) $\varepsilon_{\mathbf{X}} \sim \mathcal{N}(\mathbf{0}_{p_X}; \Psi_X)$. (4) $\varepsilon_{\mathbf{M}} \sim \mathcal{N}(\mathbf{0}_{p_M}; \Psi_M)$.
(5) Ψ_X is a $p_X \times p_X$ semi-positive definite matrix.
(6) Ψ_M is a $p_M \times p_M$ semi-positive definite matrix.
(7) $\varepsilon_Y \sim \mathcal{N}\big(0, \sigma_Y^2\big)$.
(8) W and C are respectively $p_X \times r$ and $p_M \times r$ semi-orthogonal matrices .
(9) A and B are respectively $p_X \times s$ and $p_M \times s$ semi-orthogonal matrices .
(10) $r < \min(p_X, p_M)$. (11) $s < \min(p_X, p_M)$.



(a) (b)

Figure 6.2: (a) Another latent variable model for mediation analysis, inspired by our extension of the PPLS model proposed by el Bouhaddani et al. (2018). (b) Graphical representation of the latent variable model for mediation analysis presented in Figure 6.2 (a). Circled nodes represent sets of latent variables.

signatures with maximal indirect effects, is very instinctive, and might be extended to our specific setting. Eventually, confounders will have to be considered too. Moreover, extensions to time-varying mediation analysis models (with the time-varying exposures, mediators, and confounders) could be considered.

Lastly, in the project on high-dimensional mediation analysis, we considered an L_1-penalized EM algorithm to encourage sparsity in some of the model parameters. Under simple regression models, various extensions of the lasso have been proposed. In particular, the adaptive lasso uses a weighted version of the L_1-norm for the penalty term, and was shown to generally outperform the lasso for appropriate choices of the weights. On the other hand, irrespective of the particular choice of the weights, an appropriate value for the tuning parameter has to be selected, and a popular strategy for its calibration relies on K-fold cross-validation. In Chapter 5, we considered the simple setting of linear regression models, and empirically showed that the K-fold cross-validation is not suitable for the calibration of the tuning parameter in the adaptive lasso, for several standard choices of the weights. We further proposed a simple alternative to rectify the defect, and empirically showed that it outperformed the usual K-fold cross-validation. Then, this method could be extended for the calibration of the tuning parameter in the adaptive lasso version of our model for high dimensional mediation analysis.

Bibliography

Aalen, O., Røysland, K., Gran, J., Kouyos, R., and Lange, T. (2016). Can we believe the dags? a comment on the relationship between causal dags and mechanisms. Statistical Methods in Medical Research, 25(5):2294–2314. PMID: 24463886.

Abdi, H., Chin, W. W., Esposito Vinzi, V., Russolillo, G., and Trinchera, L. (2013). New Perspectives in Partial Least Squares and Related Methods. Springer Proceedings in Mathematics & Statistics.

Agudo, A., Bonet, C., Travier, N., González, C., Vineis, P., Bueno-de Mesquita, H., Trichopoulos, D., Boffetta, P., Clavel-Chapelon, F., Boutron-Ruault, M.-C., Kaaks, R., Lukanova, A., Schütze, M., Boeing, H., Tjønneland, A., Halkjær, J., Overvad, K., Dahm, C., Quirós, J., and Riboli, E. (2012). Impact of cigarette smoking on cancer risk in the european prospective investigation into cancer and nutrition study. Journal of Clinical Oncology, pages 4550–4557.

Ahrens, W. and Pigeot, I. (2005). Handbook of Epidemiology. Springer.

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. Technometrics, 16(1):125–127.

Arlot, S. (2008). V -fold cross-validation improved: V -fold penalization. arXiv preprint arXiv:0802.0566v2.

Arlot, S. (2019). Minimal penalties and the slope heuristics: a survey. arXiv preprint arXiv:1901.07277.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. Statistics surveys, 4:40–79.

Arnold, M., Charvat, H., Freisling, H., Noh, H., Adami, H.-O., Soerjomataram, I., and Weiderpass, E. (2019). Adulthood overweight and survival from breast and colorectal cancer in swedish women. Cancer Epidemiology and Prevention Biomarkers.

Arnold, M., Freisling, H., Stolzenberg-Solomon, R., Kee, F., O'Doherty, M., Ordóñez Mena, J. M., Wilsgaard, T., May, A., Bueno-de Mesquita, H., Tjønneland, A., Orfanos, P., Trichopoulou, A., Boffetta, P., Bray, F., Jenab, M., and Soerjomataram, I. (2016).

Overweight duration in older adults and cancer risk: a study of cohorts in europe and the united states. European Journal of Epidemiology, 31(9):893–904.

Assi, N., Fages, A., Vineis, P., Chadeau-Hyam, M., Stepien, M., Duarte-Salles, T., Byrnes, G., Boumaza, H., Knüppel, S., Kühn, T., Palli, D., Bamia, C., Boshuizen, H., Bonet, C., Overvad, K., Johansson, M., Travis, R., Gunter, M., Lund, E., Dossus, L., Elena-Herrmann, B., Riboli, E., Jenab, M., Viallon, V., and Ferrari, P. (2015a). A statistical framework to model the meeting-in-the-middle principle using metabolomic data: application to hepatocellular carcinoma in the EPIC study. Mutagenesis, 30(6):743–753.

Assi, N., Moskal, A., Slimani, N., Viallon, V., Chajes, V., Freisling, H., Monni, S., Knueppel, S., Förster, J., Weiderpass, E., Lujan-Barroso, L., Amiano, P., Ardanaz, E., Molina-Montes, E., Salmerón, D., Ramón, J., Dossus, L., Fournier, A., Baglietto, L., Turzanski Fortner, R., Kaaks, R., Trichopoulou, A., Bamia, C., Orfanos, P., Santucci De Magistris, M., Masala, G., Agnoli, C., Ricceri, F., Tumino, R., Bueno de Mesquita, B., Bakker, M., Peeters, P., Skeie, G., Braaten, T., Winkvist, A., Johansson, I., Khaw, K.-T., Wareham, N., Key, T., Travis, R., Schmidt, J., Merritt, M., Riboli, E., Romieu, I., and Ferrari, P. (2015b). A treelet transform analysis to relate nutrient patterns to the risk of hormonal receptor-defined breast cancer in the european prospective investigation into cancer and nutrition (epic). Public Health Nutrition, pages 1–13.

Avin, C., Shpitser, I., and Pearl, J. (2005). Identifiability of path-specific effects. volume 19, pages 357–363.

Bach, F. and Jordan, M. (2005). A probabilistic interpretation of canonical correlation analysis.

Bagnardi, V., Rota, M., Botteri, E., Tramacere, I., Islami, F., Fedirko, V., Scotti, L., Jenab, M., Turati, F., Pasquali, E., Pelucchi, C., Galeone, C., Bellocco, R., Negri, E., Corrao, G., Boffetta, P., and Vecchia, C. (2015). Alcohol consumption and site-specific cancer risk: a comprehensive dose-response meta-analysis. British Journal of Cancer, 112(3):580–593.

Balke, A. and Pearl, J. (1994). Probabilistic evaluation of counterfactual queries. In Proceedings of the 12th Conference on Artificial Intelligence, Volume 1, Menlo Park, CA,. MIT Press.

Ballout, N., Garcia, C., and Viallon, V. (2020). Sparse estimation for case-control studies with multiple disease subtypes. Biostatistics, To Appear.

Ballout, N. and Viallon, V. (2017). Structure estimation of binary graphical models on stratified data: application to the description of injury tables for victims of road accidents. Statistics in Medicine, 38(14):2680–2703.

Barker, M. and Rayens, W. (2003). Partial least squares for discrimination, journal of chemometrics. Journal of Chemometrics, 17:166 – 173.

Basilevsky, A. (1994). Statistical Factor Analysis and Related Methods. New York: Wiley.

Bradbury, K. E., Appleby, P. N., Tipper, S. J., Travis, R. C., Allen, N. E., Kvaskoff, M., Overvad, K., Tjønneland, A., Halkjær, J., Cervenka, I., et al. (2019). Circulating insulin-like growth factor i in relation to melanoma risk in the european prospective investigation into cancer and nutrition. International journal of cancer, 144(5):957–966.

Bühlmann, P. and Meier, L. (2008). Discussion: One-step sparse estimates in nonconcave penalized likelihood models. Ann. Statist., 36(4):1534–1541.

Bühlmann, P. and van De Geer, S. (2011). Statistics for high-dimensional data: methods, theory and applications. Springer Science & Business Media.

Chan, A. T., Ogino, S., Giovannucci, E. L., and Fuchs, C. S. (2011). Inflammatory markers are associated with risk of colorectal cancer and chemopreventive response to anti-inflammatory drugs. Gastroenterology, 140(3):799–808.

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. Biometrika, 95(3):759–771.

Chichignoud, M., Lederer, J., and Wainwright, M. J. (2016). A practical scheme and fast algorithm to tune the lasso with optimality guarantees. The Journal of Machine Learning Research, 17(1):8162–8181.

Chiquet, J., Mariadassou, M., and Robin, S. (2017). Variational inference for probabilistic poisson pca. The Annals of Applied Statistics, 12:2674–2698.

Chén, O., Crainiceanu, C., Ogburn, E., Caffo, B., Wager, T., and Lindquist, M. (2018). High-dimensional multivariate mediation: with application to neuroimaging data. Biostatistics, 19:121–136.

Cirulli, E., Guo, L., Swisher, C., Shah, N., Huang, L., Napier, L., Kirkness, E., Spector, T., Caskey, C., Thorens, B., Venter, J., and Telenti, A. (2019). Profound perturbation of the metabolome in obesity is associated with health risk. Cell Metabolism, 29.

Cole, S. R. and Frangakis, C. E. (2009). The consistency statement in causal inference: a definition or an assumption? Epidemiology, 20(1):3–5.

Daniel, R. M., Cousens, S., DE Stavola, B. L., Kenward, M. G., and Sterne, J. A. (2012). Methods for dealing with time-dependent confounding. Statistics in Medicine, 32:1584–1618.

Dashti, S. G., Simpson, J., Karahalios, A., Viallon, V., Moreno-Betancur, M., Gurrin, L., Macinnis, R., Lynch, B., Baglietto, L., Morris, H., Gunter, M., Ferrari, P., Milne, R., Giles, G., and English, D. (2019). Adiposity and estrogen receptor-positive, postmenopausal breast cancer risk: Quantification of the mediating effects of fasting insulin and free estradiol. International Journal of Cancer, 146.

Dashti, S. G., Viallon, V., Simpson, J., Karahalios, A., Moreno-Betancur, M., English, D., Gunter, M., and Murphy, N. (2020). Explaining the link between adiposity and colorectal cancer risk in men and postmenopausal women in the uk biobank: A sequential causal mediation analysis. International Journal of Cancer.

Dawid, A. P. (2000). Causal inference without counterfactuals. Journal of the American Statistical Association, 95(450):407–424.

De Rubeis, V., Cotterchio, M., Smith, B. T., Griffith, L. E., Borgida, A., Gallinger, S., Cleary, S., and Anderson, L. N. (2019). Trajectories of body mass index, from adolescence to older adulthood, and pancreatic cancer risk; a population-based case–control study in ontario, canada". Cancer Causes & Control, 30(9):955–966.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 39(1):1–38.

Derkach, A., Pfeiffer, R. M., Chen, T.-H., and Sampson, J. N. (2019). High dimensional mediation analysis with latent variables. Biometrics, 75(3):745–756.

Ding, P. and VanderWeele, T. J. (2016). Sensitivity analysis without assumptions. Epidemiology (Cambridge, Mass.), 27:368–377.

Donoho, D. L. et al. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. AMS math challenges lecture, 1(2000):32.

Dossus, L., Lukanova, A., Rinaldi, S., Allen, N., Cust, A. E., Becker, S., Tjonneland, A., Hansen, L., Overvad, K., Chabbert-Buffet, N., et al. (2013). Hormonal, metabolic, and inflammatory profiles and endometrial cancer risk within the epic cohort—a factor analysis. American journal of epidemiology, 177(8):787–799.

Du, D., Bruno, R., Blizzard, L., Venn, A., Dwyer, T., Smith, K. J., Magnussen, C. G., and Gall, S. (2020). The metabolomic signatures of alcohol consumption in young adults. European Journal of Preventive Cardiology, 27(8):840–849. PMID: 30857428.

Durif, G., Modolo, L., Mold, J. E., Lambert-Lacroix, S., and Picard, F. (2019). Probabilistic count matrix factorization for single cell expression data analysis. Bioinformatics, 35(20):4011–4019.

el Bouhaddani, S., Uh, H.-W., Hayward, C., Jongbloed, G., and Houwing, J. (2018). Probabilistic partial least squares model: identifiability, estimation and application. Journal of Multivariate Analysis, 167:331–346.

Etievant, L. and Viallon, V. (2020). Causal inference under over-simplified longitudinal causal models. arXiv preprint arXiv:1810.01294.

Fan, A. Z., Russell, M., Stranges, S., Dorn, J., and Trevisan, M. (2008). Association of Lifetime Alcohol Drinking Trajectories with Cardiometabolic Risk. The Journal of Clinical Endocrinology & Metabolism, 93(1):154–161.

Fan, J. and Li, R. (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery. arXiv preprint math/0602133.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. Journal of statistical software, 33(1):1.

Ge, Z., Gao, F., and Song, Z. (2011). Mixture probabilistic PCR model for soft sensing of multimode processes. Chemometrics and Intelligent Laboratory Systems, 105:91–105.

Geisser, S. (1975). The predictive sample reuse method with applications. Journal of the American statistical Association, 70(350):320–328.

Geuter, S., Reynolds Losin, E., Roy, M., Atlas, L., Schmidt, L., Krishnan, A., Koban, L., Wager, T., and Lindquist, M. (2020). Multiple brain networks mediating stimulus-pain relationships in humans. Cerebral Cortex, 30(7):4204–4219.

Giacobino, C., Sardy, S., Diaz-Rodriguez, J., Hengartner, N., et al. (2017). Quantile universal threshold. Electronic Journal of Statistics, 11(2):4701–4722.

Giraud, C. (2014). Introduction to high-dimensional statistics. Chapman and Hall/CRC.

Glymour, M. and Greenland, S. (2008). Causal diagrams. In Modern epidemiology, pages 183–209. Philadelphia, 3rd ed. lippincott williams & wilkins edition.

Gross, S. M. and Tibshirani, R. (2016). Data shared lasso: A novel tool to discover uplift. Computational statistics & data analysis, 101:226–235.

Guan, Y. and Dy, J. (2009). Sparse probabilistic principal component analysis. Journal of Machine Learning Research - Proceedings Track, 5:185–192.

Harada, S., Takebayashi, T., Kurihara, A., Akiyama, M., Suzuki, A., Hatakeyama, Y., Sugiyama, D., Kuwabara, K., Takeuchi, A., Okamura, T., Nishiwaki, Y., Tanaka, T., Hirayama, A., Sugimoto, M., Soga, T., and Tomita, M. (2016). Metabolomic profiling reveals novel biomarkers of alcohol intake and alcohol-induced liver injury in community-dwelling men. Environmental health and preventive medicine, 21:18–26.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). Statistical learning with sparsity: the lasso and generalizations. Chapman and Hall/CRC.

Hernán, M. A. (2016). Does water kill? a call for less casual causal inferences. Annals of Epidemiology, 26(10):674–680.

Hernán, M. A., Alonso, A., Logan, R., Grodstein, F., Michels, K. B., Stampfer, M. J., Willett, W. C., Manson, J. E., and Robins, J. M. (2008). Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. Epidemiology (Cambridge, Mass.), 19(6):766.

Hernán, M. A. and Robins, J. M. (2020). Causal Inference: What If. Boca Raton: Chapman & Hall/CRC. forthcoming.

Hernán, M. A. and Taubman, S. L. (2008). Does obesity shorten life? The importance of well-defined interventions to answer causal questions. International Journal of Obesity, 32:S8–S14.

Hernán, M. A. and VanderWeele, T. J. (2011). Compound treatments and transportability of causal inference. Epidemiology (Cambridge, Mass.), 22(3):368.

His, M., Viallon, V., Dossus, L., Gicquiau, A., Achaintre, D., Scalbert, A., Ferrari, P., Romieu, I., Onland-Moret, N., Weiderpass, E., Dahm, C., Overvad, K., Olsen, A., Tjønneland, A., Fournier, A., Rothwell, J., Severi, G., Kühn, T., Fortner, R., and Rinaldi, S. (2019). Prospective analysis of circulating metabolites and breast cancer in epic. BMC Medicine, 17:178.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1):55–67.

Höskuldsson, A. (1988). PLS regression methods. Journal of Chemometrics, 2(3):211–228.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24.

Hotelling, H. (1936). Relations between two sets of variables. Biometrika, 28(3-4):321–377.

Howlader, N., Noone, A., Krapcho, M., Miller, D., Brest, A., Yu, M., Ruhl, J., Tatalovich, Z., Mariotto, A., Lewis, D., Chen, H., Feuer, E., and Kronin, K. (2020). Seer cancer statistics review, 1975-2016. Technical report, National Cancer Institute.

Huang, J., Horowitz, J. L., Ma, S., et al. (2008a). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. The Annals of Statistics, 36(2):587–613.

Huang, J., Ma, S., and Zhang, C.-H. (2008b). Adaptive lasso for sparse high-dimensional regression models. Statistica Sinica, pages 1603–1618.

Huang, Y. and Valtorta, M. (2006). Identifiability in causal bayesian networks: A sound and complete algorithm.

Jacob, L., Obozinski, G., and Vert, J.-P. (2009). Group lasso with overlap and graph lasso. In Proceedings of the 26th annual international conference on machine learning, pages 433–440.

Jolliffe, I. (2002). Principal component analysis. 2nd ed. Springer.

Jolliffe, I. T. and Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065):20150202.

Jöreskog, K. and Wold, H. (1982). Soft modeling: The basic design and some extensions. In Joreskog, K. and Wold, H., editors, Systems under indirect observation. Causality, structure, prediction. (Conference held October 18-20, 1979, at Cartigny near Geneva). Part II, pages 1–54.

Khandekar, M. J., Cohen, P., and Spiegelman, B. M. (2011). Molecular mechanisms of cancer development in obesity. Nature Reviews Cancer, 11(12):886.

Kim, Y., Han, B., and group, K. (2017). Cohort profile: The korean genome and epidemiology study (koges) consortium. International Journal of Epidemiology, e20:1–10.

Krishnan, A., Williams, L., McIntosh, A., and Abdi, H. (2011). Partial least squares (PLS) methods for neuroimaging: A tutorial and review. NeuroImage, 56:455–75.

Krumsiek, J., Suhre, K., Illig, T., Adamski, J., and Theis, F. (2011). Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. BMC systems biology, 5:21.

Kunzmann, A. T., Coleman, H. G., Huang, W.-Y., and Berndt, S. I. (2018). The association of lifetime alcohol use with mortality and cancer risk in older adults: A cohort study. PLOS Medicine, 15:1–18.

Langenau, J., Boeing, H., Bergmann, M. M., Nöthlings, U., and Oluwagbemigun, K. (2019). The association between alcohol consumption and serum metabolites and the modifying effect of smoking. Nutrients, 11(10).

Lauby-Secretan, B., Scoccianti, C., Loomis, D., Grosse, Y., Bianchini, F., and Straif, K. (2016). Body fatness and cancer - viewpoint of the iarc working group. New England Journal of Medicine, 375(8):794–798.

Lauritzen, S. L. (1996). Graphical models. Oxford University Press.

Li, S., Nyagilo, J., Dave, D., Wang, W., Zhang, B., and Gao, J. (2015). Probabilistic partial least squares regression for quantitative analysis of raman spectra. International Journal of Data Mining and Bioinformatics, 11:223–243.

Loh, W., Moerkerke, B., Loeys, T., and Vansteelandt, S. (2020). Non-linear mediation analysis with high-dimensional mediators whose causal structure is unknown. arXiv preprint arXiv:2001.07147.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Statistics in medicine, 23(19):2937–2960.

Maxwell, S. E. and Cole, D. A. (2007). Bias in cross-sectional analyses of longitudinal mediation. Psychological Methods, 12:23–44.

Maxwell, S. E., Cole, D. A., and Mitchell, M. A. (2011). Bias in cross-sectional analyses of longitudinal mediation: Partial and complete mediation under an autoregressive model. Multivariate Behavioral Research, 46(11):816–841.

Ollier, E. and Viallon, V. (2017). Regression modeling on stratified data with the lasso. Biometrika, 104(1):84–96.

Park, C., Wang, M., and Mo, E. (2017). Probabilistic penalized principal component analysis. Communications for Statistical Applications and Methods, 24:143–154.

Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Pearl, J. (1993). Comment: Graphical models, causality and intervention. Statistical Science, 8(3):266–269.

Pearl, J. (1995). Causal diagrams for empirical research. Biometrika, 82(4):669–688.

Pearl, J. (2000). Causality: models, reasoning, and inference. Cambridge University Press.

Pearl, J. (2001). Direct and Indirect Effects. In Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence, pages 411–420, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Pearl, J. (2010). On the consistency rule in causal inference: axiom, definition, assumption, or theorem? Epidemiology, 21(6):872–875.

Pearl, J., Glymour, M., and Jewell, N. P. (2016). Causal inference in statistics: a primer. John Wiley & Sons.

Pearl, J. and Robins, J. M. (1995). Probabilistic evaluation of sequential plans from causal models with hidden variables. ArXiv preprint arXiv:1302.4977.

Petersen, M. L. (2011). Compound treatments, transportability, and the structural causal model: the power and simplicity of causal graphs. Epidemiology, 22(3):378–381.

Petersen, M. L. and van der Laan, M. J. (2014). Causal models and learning from data: integrating causal modeling and statistical estimation. Epidemiology (Cambridge, Mass.), 25(3):418.

Petimar, J., Tabung, F. K., Valeri, L., Rosner, B., Chan, A. T., Smith-Warner, S. A., and Giovannucci, E. L. (2018). Mediation of associations between adiposity and colorectal cancer risk by inflammatory and metabolic biomarkers. International journal of cancer.

Platt, A., Sloan, F., and Costanzo, P. (2010). Alcohol-consumption trajectories and associated characteristics among adults older than age 50*. Journal of studies on alcohol and drugs, 71:169–79.

Renehan, A. G., Zwahlen, M., and Egger, M. (2015). Adiposity and cancer risk: new mechanistic insights from epidemiology. Nature Reviews Cancer, pages 484–498.

Riboli, E., Hunt, K., Slimani, N., Ferrari, P., Norat, T., Fahey, M., Charrondiere, U., Hémon, B., Casagrande, C., Vignat, J., Overvad, K., Tjønneland, A., Clavel-Chapelon, F., Thiébaut, A., Wahrendorf, J., Boeing, H., Trichopoulos, D., Trichopoulou, A., Vineis, P., and Saracci, R. (2002). European prospective investigation into cancer and nutrition (epic): study populations and data collection. Public health nutrition, 5(6b):1113–1124.

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):1393–1512.

Robins, J. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. Epidemiology, 3(2):143—155.

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

Rosipal, R. and Krämer, N. (2006). Overview and recent advances in partial least squares. In Saunders, C., Grobelnik, M., Gunn, S., and Shawe-Taylor, J., editors, Subspace, Latent Structure and Feature Selection, pages 34–51. Springer Berlin Heidelberg.

Rothman, K. J. and Greenland, S. (2005). Causation and causal inference in epidemiology. American Journal of Public Health, 95(S1):S144–S150.

Rothman, K. J., Greenland, S., and Lash, T. L. (2008). Modern Epidemiology. Lippincott Williams & Wilkins.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of educational Psychology, 66(5):688.

Rubin, D. B. (1980). Comment on: "randomization analysis of experimental data: The fisher randomization test " by D. Basu. Journal of the American Statistical Association, 75(371):591–593.

Sampson, J. N., Boca, S. M., Moore, S. C., and Heller, R. (2018). FWER and FDR control when testing multiple mediators. Bioinformatics, 34(14):2418–2424.

Sampson, P. D., Streissguth, A. P., Barr, H. M., and Bookstein, F. L. (1989). Neurobehavioral effects of prenatal alcohol: Part ii. partial least squares analysis. Neurotoxicology and Teratology, 11(5):477 – 491.

Schairer, C., Fuhrman, B. J., Boyd-Morin, J., Genkinger, J. M., Gail, M. H., Hoover, R. N., and Ziegler, R. G. (2016). Quantifying the role of circulating unconjugated estradiol in mediating the body mass index–breast cancer association. Cancer Epidemiology and Prevention Biomarkers, 25(1):105–113.

Shpitser, I. and Pearl, J. (2006). Identification of joint interventional distributions in recursive semi-markovian causal models. volume 2.

Siegel, J. (2019). Accelerated optimization with orthogonality constraints.

Slimani, N., Kaaks, R., Ferrari, P., Casagrande, C., Clavel-Chapelon, F., Lotze, G., Kroke, A., Trichopoulos, D., Trichopoulou, A., Lauria, C., Bellegotti, M., Ocke, M., Peeters, P., Engeset, D., Lund, E., Agudo, A., Larrañaga, N., Mattisson, I., Aronsson, C., and Riboli, E. (2002). European prospective investigation into cancer and nutrition (epic) calibration study: rationale, design and population characteristics. Public health nutrition, 5(6b):1125–1145.

Smilde, A., Bro, R., and Geladi, P. (2004). Multi Way Analysis — Applications in Chemical Sciences. New York: Wiley.

Spirtes, P., Glymour, C., and Scheines, R. (1993). Causation, Prediction, and Search. Springer-Verlag New York.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society, Series B, 36:111–147.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS medicine, 12(3):e1001779.

Tian, J. and Pearl, J. (2002). A general identification condition for causal effects. Proceedings of the Eighteenth National Conference on Artificial Intelligence.

Tian, J. and Pearl, J. (2003). On the identification of causal effects. technical report.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1):267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(1):91–108.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61(3):611–622.

van der Laan, M., Haight, T., and Tager, I. (2005). Discussion: Hypothetical interventions to define causal effects: afterthought or prerequisite. The American Journal of Epidemiology, 162:382–88.

Vandenbroucke, J. P., Broadbent, A., and Pearce, N. (2016). Causality and causal inference in epidemiology: the need for a pluralistic approach. International Journal of Epidemiology, 45(6):1776–1786.

VanderWeele, T. J. (2015). Explanation in Causal Inference - Methods for Mediation and Interaction. Oxford University Press.

VanderWeele, T. J. and Arah, O. (2011). Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. Epidemiology (Cambridge, Mass.), 22:42–52.

VanderWeele, T. J. and Hernán, M. A. (2013). Causal inference under multiple versions of treatment. Journal of causal inference, 1(1):1–20.

VanderWeele, T. J. and Tchetgen Tchetgen, E. (2017). Mediation analysis with time-varying exposures and mediators. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 79:917–938.

VanderWeele, T. J., Vansteelandt, S., and Robins, J. (2014). Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. Epidemiology, 25:300–6.

Verma, T. and Pearl, J. (1988). Causal networks: Semantics and expressiveness. Proceedings of the Fourth Workshop on Uncertainty in Artificial Intelligence.

Viallon, V., Lambert-Lacroix, S., Hoefling, H., and Picard, F. (2016). On the robustness of the generalized fused lasso to prior specifications. Statistics and Computing, 26(1):285–301.

Wegelin, J. A. (2000). A survey of partial least squares (PLS) methods, with emphasis on the two-block case. Technical report, University of Washington, Department of Statistics.

Wen, Z. and Yin, W. (2010). A feasible method for optimization with orthogonality constraints. Mathematical Programming, 142:397–434.

Wild, C., Weiderpass, E., and Stewart, B. (2020). World cancer report: Cancer research for cancer prevention. Technical report, World Cancer Report: Cancer Research for Cancer Prevention.

Wold, H. (1985). Partial least squares. In Kotz, S. and Johnson, N., editors, Encyclopedia of Statistical Sciences. Volume 6, pages 581–591.

Yang, Y., Dugu, P.-A., Lynch, B. M., Hodge, A. M., Karahalios, A., MacInnis, R. J., Milne, R. L., Giles, G. G., and English, D. R. (2019). Trajectories of body mass index in adulthood and all-cause and cause-specific mortality in the melbourne collaborative cohort study. BMJ Open, 9(8).

Zeng, J., Liu, K., Huang, W., and Liang, J. (2017). Sparse probabilistic principal component analysis model for plant-wide process monitoring. Korean Journal of Chemical Engineering, 34:1–12.

Zhang, J., Jeng, X. J., and Liu, H. (2008). Some two-step procedures for variable selection in high-dimensional linear regression. arXiv preprint arXiv:0810.1644.

Zhao, Y. and Luo, X. (2016). Pathway lasso: Estimate and select sparse mediation pathways with high dimensional mediators. arXiv preprint arXiv:1603.07749.

Zheng, J., Song, Z., and Ge, Z. (2016). Probabilistic learning of partial least squares regression model: Theory and industrial applications. Chemometrics and Intelligent Laboratory Systems, 158:80 – 90.

Zheng, R., Du, M., Zhang, B., Xin, J., Chu, H., Ni, M., Zhang, Z., Gu, D., and Wang, M. (2018). Body mass index (bmi) trajectories and risk of colorectal cancer in the plco cohort. In British Journal of Cancer.

Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American statistical association, 101(476):1418–1429.

Appendix A

Tools and principles of causal inference

The first objective of this Appendix is to briefly describe some of the challenges faced notably in epidemiology to answer etiologic questions. Some basic tools and principles devised for this purpose in the causal inference literature, and used to derive some of the methodological results presented in this thesis, are then reviewed for both time-fixed and longitudinal configurations, in Section A.1 and Section A.3. A similar review is presented for mediation analysis in Section A.2, for time-fixed configurations.

A.1 Causal inference

A.1.1 Association and causation

One main interest in Statistics is to establish association between variables. On the other hand, epidemiology is more concerned about causal interpretation, as its aim is to devise strategies to improve health. Consider for instance the example where a biomarker is found to be associated with the occurrence of a certain disease using "conventional" statistical tools. If the association is strong enough, this biomarker could be used, e.g. for diagnostic or prognostic purposes. However, epidemiologists usually want to go one step further and assess whether the biomarker is a cause of this disease: if so, intervening on the biomarker level could help prevent or cure the disease.

However, if two variables X and Y are associated, for example biomarker and disease occurrence, respectively, it does not imply that X is a cause of Y: Y could be a cause of X, or X and Y could share a common cause, etc. In other words, association does not imply causation (Hernán and Robins, 2020). In particular, in the situation depicted in Figure A.1 (c), the relationship between X and Y is not causal at all, and a spurious association exists between these two variables because of the presence of a shared cause W. Such common causes are usually referred to as confounders, as they confound and pollute the relationship between two other variables. We will come back to the use of graphs to depict such situations, as proposed in Figure A.1.

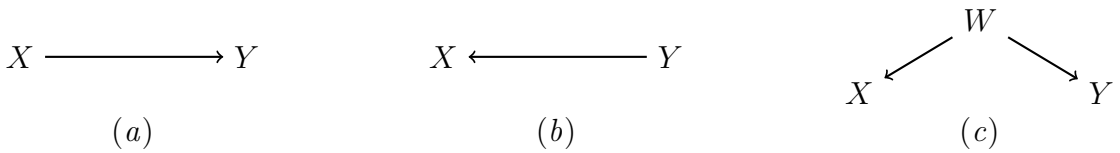


Figure A.1: (a) X is a cause of Y. (b) Y is a cause of X. (c) X and Y share a common cause, W.

Then, a natural question is how to infer causation from association measures. When using interventional data, such as in randomized clinical trials, established correlations can usually be read causally. Indeed, imagine that a randomized clinical trial is carried out to study the causal effect of a blood biomarker: a manipulation of the biomarker is randomly performed on half of the individuals of a sample of patients with a disease, who constitute the treatment group, and the other half of the sample constitutes the control group. Then the difference between the recovery rate in the two groups is usually interpreted as the causal effect of the biomarker. Indeed, in this type of experimental study, an intervention is performed and the data then directly reflect the consequences of the intervention. In particular, the randomized attribution of the exposure preserves the estimation of causal effects from confounding bias: thanks to the randomization, the two groups (treated and non-treated) are "exchangeable", in the sense that the two groups (are supposed to) have the same characteristics before treatment assignment. Such interventional data have to be contrasted with observational data, obtained for instance from a cohort study, and where no intervention is performed. For many reasons, e.g. costs or ethical reasons, epidemiologists usually have access to observational data only.

The statistical investigation of causal effects from observational data faces a key challenge, as conventional statistical models and tools only target the statistical association between variables. Yet, epidemiologists need to go beyond the results of association analyses to answer etiologic questions from observational data. Then, in order to distinguish the three different situations displayed in Figure A.1, new tools are needed. These tools have been recently devised in the causal inference literature. As we will see below, it is important to keep in mind that causal inference using observational data will invariably rely on assumptions, or prior knowledge, on the causal "system" under consideration.

A.1.2 The main tools of causal inference

A substantial literature has recently emerged, and new tools have been developed to formally define causal effects. Statistical methods were further developed to estimate these causal effects from observational data. Devised jointly by the statistics, mathematics and informatics communities, some of the tools and principles of the causal inference literature are derived from the probabilistic graphical model theory; this is notably the case for Structural Causal Models, which combine graphical causal models and structural equation

models to specify knowledge and assumptions on the causal system of interest (Pearl, 1995, 2000). In particular, graphical causal models rely on the use of Directed Acyclic Graphs, and the visual aspect they offer is particularly convenient to capture and exploit the information. The causal inference literature further proposes to use counterfactual variables to offer a formal definition of the causal effect of an exposure on an outcome (Pearl, 1995, Rubin, 1974). Moreover, a number of sets of sufficient conditions have been established, that ensure that such causal quantity of interest can be "identified", or in other words, be expressed in terms of observable quantities only.

In this Section, the progression chosen to present these tools and principles is largely inspired by Pearl et al. (2016).

Using graphical causal models and structural equation models to specify the assumptions on the causal system of interest

In order to infer causation from association measures, prior knowledge and/or assumptions on the causal system of interest need to be used. They typically concern the possible relationships between the variables in the causal system, and are *de facto* related to the structure of the graph that can be used to summarize them.

A graphical causal model (Lauritzen, 1996, Pearl, 1988) offers a visual approach to clearly describe these assumptions. The exposure of interest, outcome, but also possible confounders are referred to as "endogenous" variables, and each of them has to be represented by a node. Each endogenous variable can also be caused by at least one "exogenous" variable; exogenous variables correspond to "external" factors, and are not necessarily drawn in the graph. The possible relationships between the endogenous variables are then translated via directed edges. In particular, an edge directed from node X to node Y represents a causal dependence between these two variables: it means that the value taken by Y may depend on the value taken by X, and not the other way round. Then, the assumptions embodied in the causal graph lie in the absence of edge, and in the direction of the edges present in the graph. For instance in the causal graph of Figure A.2, X is a possible cause of M and Y, but Y cannot directly cause X because there is no directed edge from Y to X. It is standard to assume that variables cannot have an effect, either directly or through other variables, on themselves, so that the graphs are Directed Acyclic Graphs (DAGs). Possible "feedback" on variables can be considered using DAGs by ackowledging the time-varying nature of variables; we will consider such settings in Section A.3.

Relationships among the variables in the DAG are usually described using the vocabulary borrowed from the graph theory: if an edge is directed from node X to node Y, then X is said to be a "parent" of Y and Y is said to be a "child" of X. In addition, if Z is a parent of X, and X is a parent of Y, Z is said to be an "ancestor" of Y and Y is said to be a "descendant" of Z. For example in Figure A.2, M is parent of Y and a child of W_1 and

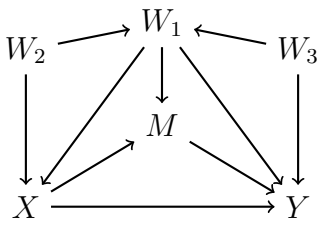


Figure A.2: Example of graphical causal model where X is a possible cause of M and Y, and is possibly caused by W_1 and W_2. M is also a possible cause of Y and is possibly caused by W_1, while W_1 is a possible cause of Y and is possibly caused by W_2 and W_3. Finally, Y is also possibly caused by W_3.

X, and is further an descendant of W_2 and W_3. Finally, note that an exogenous variable cannot be a descendant of any endogenous variable or any other exogenous variable.

A structural equation model (SEM) provides a complementary approach to formally specify the assumptions on the causal system of interest. For this purpose, Pearl (1995, 2000) proposes to use a Structural Causal Model (SCM), which combines a graphical causal model and a system of structural equations. In the SEM, each endogenous variable is defined as the output of some function, whose inputs are some exogenous and possibly other endogenous variables; then just as the DAG, the SEM specifies our assumptions on the structure of the causal system. When the functions intervening in the structural equations are completely specified, the SEM brings more information than the graphical model, as it clearly indicates how parents cause each of their child. However, these functions do not have to be completely specified; the only requirement is to specify which endogenous and exogenous variables are possible inputs of the functions. Then, in the case of a non-parametric SEM, the system of structural equations and the DAG contain exactly the same information. Consider for instance the SCM given in Figure A.3 (a); the system of structural equations indicates that variable X is defined only from the value of variable U_X, and Y is defined from the value of both X and U_Y. One the other hand, neither U_X nor U_Y are defined from a specified function; they are exogenous variables. Regarding the graphical model, X as a unique parent (U_X), Y has two parents (X and U_Y), and U_X and U_Y have no parent. Then, the causal graph can be drawn given a system of structural equations , and *vice versa*. Note however that in the causal graph of Figure A.3 (a), the absence of directed edges between U_X and U_Y implicitly suggests that the variables are independent. Causal models where exogenous variables are independent of each other are called Markovian (Pearl, 2000); we assume, unless otherwise stated, that this is the case in the forthcoming causal models.

Figure A.3: (a) Example of SCM, where X is a potential cause of Y. X and Y are also potentially cause by exogenous variables U_X and U_Y, respectively. (b) Causal diagram and system of structural equations representing the causal system of Figure A.3 (a) in the counterfactual world where the exposure X would have been set to value x.

Using specific configurations of edges and nodes in graphical causal models to read independence between variables

In a DAG, paths between two nodes consist in specific configurations of edges, which may be conditionally or unconditionally blocked. As in a graphical causal model each node corresponds to a variable, independence relationships between variables will be implied by the structure of the DAG, provided that the probability distribution of the variables is compatible with the structure of the DAG. In this Section, we quickly present the different types of configurations of edges and nodes which may appear in a DAG, as well as how such paths can be blocked (Pearl, 1988). Then, we present the d-separation criterion proposed by Verma and Pearl (1988), which allows sets of (conditional) independencies among the variables composing a DAG to be deduced from its structure. It will be particularly useful for the investigation of the identifiability of causal effects, for example when using the twin networks that will be introduced below.

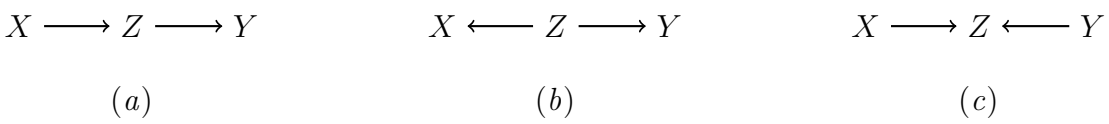


Figure A.4: Graphical causal models depicting: (a) A chain. (b) A fork. (c) A collider. For readability, exogenous variables U_X, U_Z and U_Y have been dropped in the graphical representations.

In DAGs, restricting our attention to any three "adjacent" nodes, only three different types of configurations exist: "chains", "forks" and "colliders"; see Figure A.4 for a graphical representation of these configurations. Then, each path in a DAG can be seen as a succession of chains, forks and/or colliders.

Definition 1. *(d-separation). A path p in a graph $\mathcal{G}$ is blocked by a set of nodes Z if and only if:*

- *p contains a chain $(A \to B \to C)$ or a fork $(A \leftarrow B \to C)$ such that B is in Z,*

  or

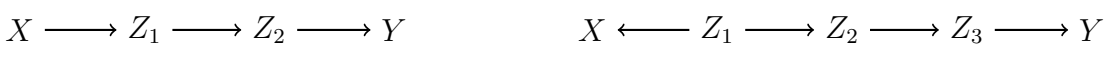
- *p contains a collider $(A \to B \leftarrow C)$ such that neither B nor any of its descendant is in Z.*

The set Z is said to d-separate X and Y in $\mathcal{G}$ if and only if Z blocks every path between X and Y in $\mathcal{G}$.

Note that d-separation may be conditional or unconditional, in the particular case where Z is the empty set. On the other hand, if two nodes are not d-separated by Z, they are said to be d-connected, conditional on Z (Pearl et al., 2016). Consider for instance the graphical causal model given in Figure A.2, and more particularly the relationship between X and Y. There exist several paths between X and Y, notably one "open" (not blocked) path, $(X \to Y)$, because of the directed edge from X to Y. Regarding the other paths:

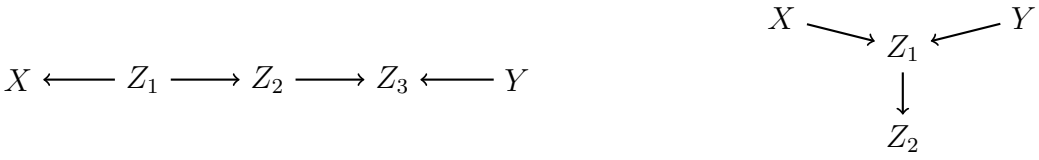
- Path p_1 $(X \to M \to Y)$ is blocked by $\{M\}$.
- Path p_2 $(X \leftarrow W_1 \to Y)$ is blocked by $\{W_1\}$.
- Path p_3 $(X \to M \leftarrow W_1 \to Y)$ is blocked by the empty set, because the collider node M is on the path. Then p_3 is not blocked by $\{M\}$, but is blocked by $\{W_1\}$ or $\{M, W_1\}$.
- Path p_4 $(X \leftarrow W_1 \to M \to Y)$ is blocked by $\{M\}$, $\{W_1\}$ or $\{M, W_1\}$.
- Path p_5 $(X \leftarrow W_2 \to W_1 \to M \to Y)$ is blocked by $\{W_2\}$, $\{W_1\}$, $\{M\}$, $\{M, W_1\}$, $\{M, W_2\}$ or $\{M, W_1, W_2\}$.
- Path p_6 $(X \to M \leftarrow W_1 \leftarrow W_3 \to Y)$ is blocked by the empty set, $\{W_1\}$, $\{W_3\}$, $\{M, W_1\}$, $\{M, W_3\}$ or $\{M, W_1, W_3\}$.
- Path p_7 $(X \leftarrow W_2 \to W_1 \to Y)$ is blocked by $\{W_1\}$, $\{W_2\}$ or $\{W_1, W_2\}$.
- Path p_8 $(X \leftarrow W_1 \leftarrow W_3 \to Y)$ is blocked by $\{W_1\}$, $\{W_3\}$ or $\{W_1, W_3\}$.
- Path p_9 $(X \leftarrow W_2 \to W_1 \leftarrow W_2 \to Y)$ is blocked by the empty set, $\{W_2\}$, $\{W_3\}$, $\{W_1, W_2\}$, $\{W_1, W_3\}$ or $\{W_1, W_2, W_3\}$.

As mentioned above, there is a connection between blocked paths or d-separation of nodes in a DAG and independence relationships among the associated variables. Indeed, an open path between two nodes in a graphical causal model can be seen as flow of dependency between the associated variables. More precisely, in the chain and fork configurations given in Figure A.4 (a) and Figure A.4 (b), the path between X and Y is open: X and



(a) (b)

Figure A.5: Examples of graphical causal models where: (a) The unique path from X to Y is composed only of chains going in the same direction. (b) Z_1, the common cause of X and Y, is on the unique path between them. For readability, exogenous variables have been dropped in the graphical representation.



(a) (b)

Figure A.6: Graphical causal models where: (a) The unique path from X to Y contains a collider, Z_3. (b) The unique path from X to Y contains a collider, Z_1. Z_2 is a child of Z_1. For readability, exogenous variables have been dropped in the graphical representation.

Y are d-connected, and then X and Y are possibly dependent. On the other hand, it is possible to block the path between X and Y by conditioning on the node at its center, and then X and Y are conditionally independent given Z. Then, if there is a unique path between X and Y, and if it is only composed of chains going in the same direction, X and Y are conditionally independent given every subset of nodes of this path. This is for instance the case in Figure A.5 (a), where $\{Z_1\}$, $\{Z_2\}$ and $\{Z_1, Z_2\}$ block the path between X and Y. In the same way, if there is a unique path between X and Y which contains their common cause, then X and Y are conditionally independent given their common cause; see for instance Figure A.5 (b), where X and Y are conditionally independent given Z_1. Lastly, in the collider configuration given in Figure A.4 (c), the path between X and Y is unconditionally blocked: X and Y are independent. In particular under such configuration, conditioning on the central node Z, which is also called a collider, opens the path between X ad Y: X and Y are d-connected conditional on Z, and then X and Y may be dependent conditional on Z. The intuition is maybe less obvious than for chains and forks, but may be seen in the following way: the value of Z depends on both X and Y; then for a given value of Z, changes in X must be compensated by Y, and *vice versa*. Finally, if there is a unique path between X and Y, and if it contains a collider, then X and Y are independent, but (possibly) conditionally dependent given the collider or any of its descendants. Consider for instance the causal system depicted in Figure A.6 (a); X and Y are marginally independent, but conditioning on Z_3 could induce dependence between X and Y. Then in Figure A.6 (b), X and Y are marginally independent as well, but conditioning on $\{Z_1\}$, $\{Z_2\}$ or $\{Z_1, Z_2\}$ could induce dependence between X and Y.

The following criterion, called d-separation criterion (Verma and Pearl, 1988), summa-

rizes the (conditional) independence relationships induced by the structure of the DAG.

Theorem 5. *(d-separation criterion) If Z d-separates X and Y in a graph $\mathcal{G}$, then X and Y are conditionally independent given Z, for any probability distribution $\mathbb{P}$ that is compatible with the structure of $\mathcal{G}$.*

We will use the notations $(X \perp\!\!\!\perp Y \mid Z)_{\mathcal{G}}$ when Z d-separates X and Y in a graph $\mathcal{G}$, and $(X \perp\!\!\!\perp Y \mid Z)_{\mathbb{P}}$ when variables X and Y are conditionally independent given Z under the probability distribution $\mathbb{P}$. Then Theorem 5 states that $(X \perp\!\!\!\perp Y \mid Z)_{\mathcal{G}} \Rightarrow (X \perp\!\!\!\perp Y \mid Z)_{\mathbb{P}}$ for any probability distribution $\mathbb{P}$ which is compatible with the structure of $\mathcal{G}$. Note on the other hand that $(X \not\perp\!\!\!\perp Y \mid Z)_{\mathcal{G}} \not\Rightarrow (X \not\perp\!\!\!\perp Y \mid Z)_{\mathbb{P}}$ for any probability distribution $\mathbb{P}$ which is compatible with the structure of $\mathcal{G}$. Indeed, if X and Y are d-connected in the DAG, the associated variables are possibly dependent, but not necessarily; Pearl et al. (2016) say that the variables are "likely" dependent. More precisely, these variables will be independent for almost every distribution compatible with the DAG, but exceptions may arise. Consider for instance the case where variables X, M and Y are defined from the following Gaussian linear models: $X \sim \mathcal{N}(0,1)$, $M = \alpha X + U_M$, $Y = \beta X + \gamma M + U_Y$, with $U_Y \sim \mathcal{N}(0,1)$ and $U_M \sim \mathcal{N}(0,1)$. In particular, this distribution of variables X, M and Y is compatible with the DAG $\mathcal{G}$ given in Figure A.7. On the one hand, $(X \not\perp\!\!\!\perp Y)_{\mathcal{G}}$. On the other hand, as $Y = (\beta + \gamma\alpha)X + \gamma U_M + U_Y$, $(X \perp\!\!\!\perp Y)_{\mathbb{P}}$ if and only if $\beta = -\gamma\alpha$. This distribution is then said to be unfaithful to $\mathcal{G}$ (Spirtes et al., 1993). In order to exclude such pathological cases, the faithfulness assumption or stability assumption (Pearl, 2000) can be made; more precisely it assumes that independence relationships between variables only arise from the structure of the DAG, and are then invariant to changes of parameters in the model.

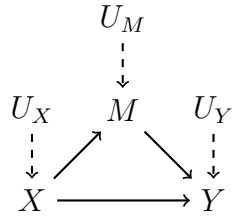


Figure A.7: Example of graphical causal model where two paths are directed from X to Y: one is simply composed of the directed edge from X to Y, and the second one is composed of a chain with M at its center.

Finally note that in this Appendix, we assume that the observable data are generated under the causal model that is (partly) specified by the considered graphical causal model. In other words, the distribution of the variables for which we have observational data is compatible with the considered DAG.

Using counterfactual variables to formalize causal effects

Let us now review the information in our possession for our causal quest. First, our assumptions on and/or knowledge of the structure of the causal system of interest can be summarized in the form of a graphical causal model and/or a SEM, from which a number of properties can be deduced for the joint distribution of the endogenous variables and

possibly available in the observational data. However, this concerns the "real world" only, while we would like to know what would have happened if an intervention had been performed on the exposure of interest X, by analogy with the data obtained from a randomized clinical trial for instance. For example, in the case where the binary exposure under study is the obesity status (obese/lean), and the outcome of interest is cancer occurrence, in order to assess the causal effect of obesity on cancer occurrence we would like to compare the risk of cancer in the population, had all the individuals of the population been obese, with the risk of cancer in the "same" population, but this time had all the individuals been lean. To do so, we can represent the causal system in another world, called counterfactual world, where contrary to the fact, an intervention would have set the exposure X to a given possible value x.

The *do*-operator (Pearl et al., 2016, Spirtes et al., 1993) operationalizes such intervention, by performing a number of "transformations" on the causal model. First, under the intervention $do(X = x)$, the edges directed to X are removed in the causal graph, and X is set to x. On the other hand in the associated SEM, the structural equation where X is the output is also replaced by the value x. Then, the hypothetical intervention on the exposure X affects the distribution of each of its descendants. Notably, X is replaced by the value x in every other structural equations, while the functions remain unmodified. In addition under the intervention $do(X = x)$, any descendant of X is replaced by a counterfactual variable; see for example Figure A.3 (b), where under the hypothetical intervention $do(X = x)$, the outcome Y becomes the counterfactual variable $Y^{X=x}$ (Pearl, 1995, Rubin, 1974).

As mentioned above, the distributions of the variables in counterfactual worlds are of particular interest when addressing causal questions. Variables that "live" in these counterfactual worlds are called counterfactual variables. In particular, the counterfactual outcome $Y^{X=x}$ is the outcome variable that would have been observed in the counterfactual world following the intervention where X would have been set to x. Taking the example of a binary exposure such as the obesity status (obese/lean), there are two hypothetical interventions: $do(X = 1)$ and $do(X = 0)$, and then two counterfactual outcomes, $Y^{X=1}$ and $Y^{X=0}$. The causal effect of X on Y (e.g., cancer occurrence) can then be defined by comparing the distributions of $Y^{X=1}$ and $Y^{X=0}$. However, such counterfactual variables are latent variables; we can never observe both $Y^{X=1}$ and $Y^{X=0}$, so the comparison between $Y^{X=1}$ and $Y^{X=0}$ cannot be performed at an "individual" level. The comparison is then performed at the "population" level: for example, considering the additive scale, $\mathbb{E}(Y^{X=1} - Y^{X=0})$ defines the average causal effect of X on Y. This quantity is sometimes also called total causal effect or average treatment effect, and denoted either ATE, for "average treatment effect", or ACE, for "average causal effect". It represents the difference between cancer risk in the counterfactual world where each individual would have been obese, and the one where each individual would have been lean. An alterna-

tive notation is also possible, using the *do*-operator instead of counterfactual variables: $\mathbb{E}(Y^{X=1} - Y^{X=0}) = \mathbb{E}(Y \mid do(X = 1)) - \mathbb{E}(Y \mid do(X = 0))$. Finally, it is noteworthy that for any possible value x of X, $\mathbb{E}(Y^{X=x}) = \mathbb{E}(Y \mid do(X = x))$ usually differs from $\mathbb{E}(Y \mid X = x)$; we will come back to this point below.

Thereafter, binary exposures and categorical confounders will be considered for simplicity. Since the causal quantity of interest is defined from counterfactuals variables $Y^{X=1}$ and $Y^{X=0}$, which are unobserved, a natural question is whether it can be expressed as a function of the observations. When a causal quantity can be expressed in terms of the distribution of observable variables only, it is said to be "identifiable". For this purpose, sets of sufficient conditions have been established to ensure that ATE can be identified. The first condition in each set of identifiability conditions is the Consistency condition (Robins, 1986)

(C) $$\text{if } X = x, \text{ then } Y^{X=x} = Y.$$

In particular, this condition allows to connect counterfactual quantities to observed ones. More precisely, it states that the outcome variable observed in the real world for an individual who actually has value x for exposure X, coincides with the value it would have taken if an intervention would have been performed to set X to value x. The second condition is the Ignorability condition (Rosenbaum and Rubin, 1983)

(Ign) $$Y^{X=x} \perp\!\!\!\perp X, \text{ for any possible value } x \text{ of } X.$$

This condition assumes an independence relationship between real world and counterfactual world variables. Note that it differs from the independence relationship $Y \perp\!\!\!\perp X$, and in particular, the ignorability condition (Ign) does not imply that $Y \perp\!\!\!\perp X$. Then, under consistency and ignorability conditions

$$\begin{aligned} ATE &= \mathbb{E}\left(Y^{X=1}\right) - \mathbb{E}\left(Y^{X=0}\right), \\ &\overset{\text{(Ign)}}{=} \mathbb{E}(Y^{X=1} \mid X = 1) - \mathbb{E}(Y^{X=0} \mid X = 0), \\ &\overset{\text{(C)}}{=} \mathbb{E}(Y \mid X = 1) - \mathbb{E}(Y \mid X = 0). \end{aligned} \tag{A.1}$$

This quantity is expressed only in terms of the distributions of observed variables X and Y. Finally, if the positivity condition holds, that is if $\mathbb{P}(X = x) > 0$, for any possible value x of X, this quantity can be estimated in practice. Another possible set of identifiability conditions is based on a "conditional version" of the former, for the ignorability and positivity conditions. Under the consistency and Conditional Ignorability

conditions (Rosenbaum and Rubin, 1983)

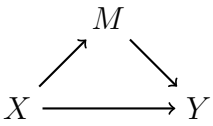
(C.Ign) $\qquad Y^{X=x} \perp\!\!\!\perp X \mid W$, for any possible value x of X.

we have

$$\begin{aligned} ATE &= \sum_w \left[\mathbb{E}\left(Y^{X=1} \mid W=w\right) - \mathbb{E}\left(Y^{X=0} \mid W=w\right)\right] \times \mathbb{P}(W=w), \\ &\overset{\text{(C.Ign)}}{=} \sum_w \left[\mathbb{E}(Y^{X=1} \mid W=w, X=1) - \mathbb{E}(Y^{X=0} \mid W=w, X=0)\right] \times \mathbb{P}(W=w), \\ &\overset{\text{(C)}}{=} \sum_w \left[\mathbb{E}(Y \mid W=w, X=1) - \mathbb{E}(Y \mid W=w, X=0)\right] \times \mathbb{P}(W=w), \end{aligned}$$

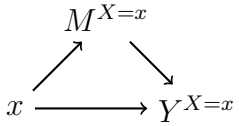
where the sum is over all possible values of W. Then, if the conditional positivity condition holds, that is if $\mathbb{P}(X=x \mid W=w) > 0$, for any possible value x of X and any possible value w of W, ATE can be estimated from observational data.

However, these identifiability conditions are not fully testable. From the observational data, it is possible to check whether the positivity or conditional positivity conditions hold, but this is not the case for the consistency, ignorability or conditional ignorability conditions, as these conditions involve counterfactual variables. Causal inference then usually relies on the structure of the causal system of interest, which summarizes other untestable but more "concrete" assumptions. In particular, the inspection of the causal model in the real and counterfactual worlds will indicate which "version" of the ignorability condition is satisfied, and then how the causal quantity of interest is identified (along with the consistency assumption). The consistency assumptions holds by construction under the SCMs; for example, under the causal system whose structures in the real world and counterfactual world following $do(X=x)$ are given in Figure A.3 (a) and Figure A.3 (b), respectively, we have $Y = f_Y(X, U_Y)$ and $Y^{X=x} = f_Y(x, U_Y)$, so that $X = x$ implies that $Y = Y^{X=x}$. The ignorability condition $\left(Y^{X=x} \perp\!\!\!\perp X\right)$ also holds under this model, as $Y^{X=x} = f_Y(x, U_Y)$, and U_Y is independent of U_X and then of X. As a result, under this configuration ATE is identified through the formula given in Equation (A.1), and to estimate the causal effect of X on Y, it is sufficient to estimate $\mathbb{E}(Y \mid X=1)$ and $\mathbb{E}(Y \mid X=0)$. Now consider the causal system whose structures in the real world and counterfactual world following $do(X=x)$ are respectively given in Figure A.8 (a) and Figure A.8 (b). $M^{X=x}$ is the M variable that would have been observed in the counterfactual world where the exposure X would have been set to value x, and in particular $M^{X=x} \perp\!\!\!\perp X$. Then $Y^{X=x} \perp\!\!\!\perp X$, and it is again sufficient to estimate $\mathbb{E}(Y \mid X=1)$ and $\mathbb{E}(Y \mid X=0)$ in order to estimate ATE. Of course the ignorability condition, in its unconditional version, does not always hold. Consider for instance the causal system whose structures in the real world and counterfactual world following $do(X=x)$ are



$$\begin{cases} X &= f_X(U_X), \\ M &= f_M(X, U_M), \\ Y &= f_Y(X, M, U_Y). \end{cases}$$

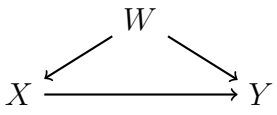
(a)



$$\begin{cases} x, & \\ M^{X=x} &= f_M(x, U_M), \\ Y^{X=x} &= f_Y(x, M^{X=x}, U_Y). \end{cases}$$

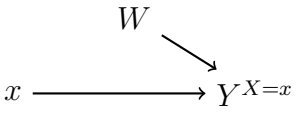
(b)

Figure A.8: (a) Example of SCM where X is a potential cause of M and Y, and where M is also a potential cause of Y. For readability, exogenous variables U_X, U_M and U_Y have been dropped in the graphical representation. (b) Causal diagram and system of structural equations representing the causal system of Figure A.8 (a) in the counterfactual world where the exposure X would have been set to value x.



$$\begin{cases} X &= f_X(W, U_X), \\ W &= f_W(U_W), \\ Y &= f_Y(X, W, U_Y). \end{cases}$$

(a)



$$\begin{cases} x, & \\ W &= f_W(U_W), \\ Y^{X=x} &= f_Y(x, W, U_Y). \end{cases}$$

(b)

Figure A.9: (a) Example of SCM where X is a potential cause of Y, and where W is a potential cause of both X and Y. For readability, exogenous variables U_X, U_Y and U_W have been dropped in the graphical representation. (b) Causal diagram and system of structural equations representing the causal system of Figure A.9 (a) in the counterfactual world where the exposure X would have been set to value x.

respectively given in Figure A.9 (a) and Figure A.9 (b); W is a potential cause of both X and Y, and is thus a confounder for the $X - Y$ relationship. In particular, $Y^{X=x} \not\perp\!\!\!\perp X$ because $Y^{X=x}$ depends on W and so does X, and then $\mathbb{E}\left(Y^{X=x}\right) \neq \mathbb{E}\left(Y \mid X = x\right)$. However $Y^{X=x} \perp\!\!\!\perp X \mid W$, and then one has to "adjust" for W: ATE is identified through $\sum_w \left[\mathbb{E}(Y \mid W = w, X = 1) - \mathbb{E}(Y \mid W = w, X = 0)\right] \times \mathbb{P}(W = w)$, where the sum is over all possible values of W.

In a nutshell, counterfactual variables allow to formally define causal quantities of interest, but also to describe sets of conditions which are sufficient to express these causal quantities in terms of the distribution of observable variables X, Y and potentially additional variables implied in the $X-Y$ relationship. Notably, in the presence of a confounder W for the $X-Y$ relationship, as in Figure A.9 (a), one has to adjust for the confounder in order to take into account the source of association between X and Y it produces. Figure A.8 (a) is another example of a causal system where an additional variable is involved in the relationship between X and Y. However here, M is not a confounder for the $X - Y$,

as even if it is a potential cause if Y, it cannot cause X. More precisely here, M is called a "mediator", as it is an intermediary factor through which the exposure X also has an "indirect" effect on Y. We have already seen that the unconditional ignorability condition $\left(Y^{X=x} \perp\!\!\!\perp X\right)$ holds under this configuration, so the marginal association between X and Y allows to recover ATE. In particular, if one were to adjust for the mediator M, it would block some of the mechanisms through which X affects Y, and then the quantity estimated in practice would not correctly assess the "total" causal effect of X on Y.

Note that one may rely on the graphical representation of the causal model or on the system of structural equations in the real and counterfactual worlds to investigate the independence relationships between counterfactual and observable variables, and then determine how causal effects are identified. However, and as previously mentioned, the visual aspect of graphical causal models is usually preferred as its reading is easier, in particular for causal systems which involve a large number of variables. Then, from now on, we will mostly focus on graphical causal models. Twin networks (Balke and Pearl, 1994, Pearl, 2000) have been proposed as a way to determine which version of the ignorability condition holds. They offer a joint representation of real and counterfactual worlds. In twin networks, the d-separation can be applied to investigate conditional independencies among variables living in different worlds. For example, Figure A.10 presents the twin network corresponding to the causal model given in Figure A.9 (a), in the real world and in the counterfactual world following $do(X = x)$. In a twin network, exogenous variables are shared between the two worlds, and so are the endogenous variables that are identical in the two worlds. One single node can be the parent of several children that correspond to variables living in different worlds. Endogenous variables that are different in the real and counterfactual worlds are "duplicated", and labeled accordingly. For instance in the causal system depicted in A.9 (a), W is identical in the real world and counterfactual world following $do(X = x)$. But, both X and Y vary when moving from one world to another, and the associated nodes have to be duplicated. Moreover, in the twin network given in Figure A.10 as well, no edge enters x. For example, in Figure A.10 it is easy to see that $\{W\}$ d-separates X and $Y^{X=x}$, and then that $Y^{X=x} \perp\!\!\!\perp X \mid W$. Another remark is that twin networks can also be useful to explore independence relationships among counterfactual variables living in different counterfactual worlds, as those involved in the mediation analysis framework; see for instance with Figure A.15 or Figure A.16 in Section A.2.1.

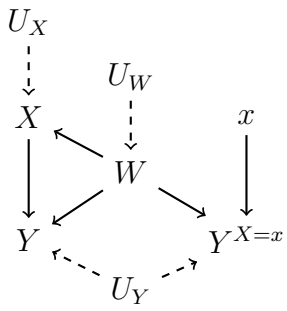


Figure A.10: Twin network representation of the causal model given in Figure A.9 (a), in the real world and in the counterfactual world following the hypothetical intervention where the exposure X would have been set to value x.

Using graphical criteria to study the identifiability of causal effects

As mentioned above, exogenous variables have been assumed to be independent of each other, and then the causal models which have been considered are all Markovian models (Pearl, 2000). It can be shown that under Markovian models, the total causal effect of any set of (endogenous) variables on any other disjoint set of (endogenous) variables is always identifiable (Pearl, 2000). Notably, we illustrated the interest of twin networks (Balke and Pearl, 1994, Pearl, 2000) to determine whether a given ignorability condition holds under a given causal model. On the other hand, several general graphical criteria, notably the back-door and front-door criteria (Pearl, 1993, 1995, 2000), have also been proposed in the literature to ensure the identifiability of total causal effects. These graphical criteria are usually preferred over twin networks, as using the later quickly becomes cumbersome when the causal system of interest involves a large number of variables. In addition, these general criteria allow to address the identifiability of causal effects in causal models with possibly unmeasured variables.

Relying on d-separation (Pearl, 1988, Verma and Pearl, 1988), the "back-door" criterion (Pearl, 1993) allows to determine whether a set of observed variables is such that the conditional ignorability assumption holds.

Theorem 6. *(Back-door criterion) A set of nodes Z satisfies the back-door criterion relative to X and Y if:*

- *No element of Z is a descendant of X,*

  and

- *Z blocks every back-door path relative to X and Y, that is all paths between X and Y with an edge directed to X and an edge directed to Y.*

If Z satisfies the back-door criterion relative to X and Y, then $Y^{X=x} \perp\!\!\!\perp X \mid Z$, for any possible value x of X.

In other words, if Z satisfies the back-door criterion relative to X and Y, Z is sufficient for the adjustment, and the causal effect of X on Y is identified as

$$ATE = \sum_z \big[\mathbb{E}(Y \mid Z = z, X = 1) - \mathbb{E}(Y \mid Z = z, X = 0)\big] \times \mathbb{P}(Z = z),$$

where the sum is over all possible values of Z. For example in the causal system depicted in Figure A.2, paths p_2, p_4, p_5, p_7, p_8 and p_9 are back-door paths relative to X and Y. Set $\{W_1\}$ does not satisfy the back-door criterion relative to X and Y as is does not block path p_9, even if it blocks the other back-door paths. However, sets $\{W_1, W_2\}$, $\{W_1, W_3\}$ or $\{W_1, W_2, W_3\}$ block every back-door path, and as neither W_1, W_2 nor W_3 is a descendant

(a) (b)

Figure A.11: (a) Example of graphical causal model where the only back-door path relative to X and Y is blocked by the collider W_2. (b) Example of graphical causal model where W is unobserved.

of X, sets $\{W_1, W_2\}$, $\{W_1, W_3\}$ or $\{W_1, W_2, W_3\}$ satisfy the back-door criterion relative to X and Y. Then for any possible value x of X, $Y^{X=x} \perp\!\!\!\perp X \mid \{W_1, W_2, W_3\}$, and

$$\begin{aligned}
\mathbb{E}\left(Y^{X=x}\right) &= \sum_{w_1}\sum_{w_2}\sum_{w_3} \mathbb{E}\left(Y^{X=x} \mid W_1 = w_1, W_2 = w_2, W_3 = w_3\right) \\
&\qquad \times \mathbb{P}\left(W_1 = w_1, W_2 = w_2, W_3 = w_3\right), \\
&\overset{\text{(C.Ign)}}{=} \sum_{w_1}\sum_{w_2}\sum_{w_3} \mathbb{E}\left(Y^{X=x} \mid X = x, W_1 = w_1, W_2 = w_2, W_3 = w_3\right) \\
&\qquad \times \mathbb{P}\left(W_1 = w_1, W_2 = w_2, W_3 = w_3\right), \\
&\overset{\text{(C)}}{=} \sum_{w_1}\sum_{w_2}\sum_{w_3} \mathbb{E}\left(Y \mid W_1 = w_1, W_2 = w_2, W_3 = w_3, X = x\right) \\
&\qquad \times \mathbb{P}\left(W_1 = w_1, W_2 = w_2, W_3 = w_3\right).
\end{aligned}$$

Finally, as $Y \perp\!\!\!\perp W_2 \mid \{X, W_1, W_3\}$ and $W_2 \perp\!\!\!\perp W_3$,

$$\begin{aligned}
ATE = \sum_{w_1}\sum_{w_2}\sum_{w_3} &\left[\mathbb{E}\left(Y \mid W_1 = w_1, W_3 = w_3, X = 1\right)\right. \\
&\left. -\mathbb{E}\left(Y \mid W_1 = w_1, W_3 = w_3, X = 0\right)\right] \\
&\times \mathbb{P}\left(W_1 = w_1 \mid W_2 = w_2, W_3 = w_3\right) \\
&\times \mathbb{P}\left(W_2 = w_2\right) \times \mathbb{P}\left(W_3 = w_3\right).
\end{aligned}$$

But on the other hand, as $\{W_1, W_2\}$ is also sufficient for the adjustment, the conditional ignorability condition $\left(Y^{X=x} \perp\!\!\!\perp X \mid \{W_1, W_2\}\right)$ also holds, and

$$\begin{aligned}
ATE = \sum_{w_1}\sum_{w_2} &\left[\mathbb{E}\left(Y \mid W_1 = w_1, W_2 = w_2, X = 1\right) - \mathbb{E}\left(Y \mid W_1 = w_1, W_2 = w_2, X = 0\right)\right] \\
&\times \mathbb{P}\left(W_1 = w_1, W_2 = w_2\right).
\end{aligned}$$

This quantity can be estimated in practice with data on variables X, Y, W_1 and W_2, and provided that some positivity assumption holds. Then under the causal model of Figure A.2, the causal effect of X of Y can be identified even if variable W_3 is unobserved. On the other hand, if the empty set satisfies the back-door criterion, the unconditional ignorability condition holds and then $\mathbb{E}\left(Y^{X=x}\right) = \mathbb{E}\left(Y \mid X = x\right)$; see Figure A.11 (a) for an example.

In addition, total causal effects can sometimes be identified and estimated in practice, even in the absence of sets of observed variables which are sufficient for the adjustment. This is notably the case when the "front-door" criterion is satisfied (Pearl, 1995). Consider for instance the causal system depicted in Figure A.11 (b), and where W, which is a confounder for the $X - Y$ relationship, is assumed to be unobserved. Under this configuration, no set of observed variable satisfies the back-door criterion relative to X and Y. However, if $\mathbb{P}(X = x, Z = z) > 0$, for any possible values x and z of X and Z, respectively, the total causal effect of X on Y is identifiable and may be estimated. Indeed, Z satisfies the front-door criterion (Pearl, 1995, 2000) relative to X and Y, as defined below.

Definition 2. *(Front-door) A set of nodes Z satisfies the front-door criterion relative to X and Y if:*

- *Z blocks every paths directed from X to Y,*

  and

- *There is no unblocked back-door path relative to X to Z,*

  and

- *All back-door paths relative to Z and Y are blocked by X.*

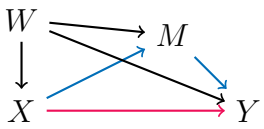
Then, the following Theorem (Pearl, 1995, 2000) can be applied.

Theorem 7. *(Front-door criterion) If Z satisfies the front-door criterion relative to X and Y, and if $\mathbb{P}(X = x, Z = z) > 0$, for any possible values x and z of X and Z, respectively, then the total causal effect of X on Y is identifiable, and is obtained with the formula*

$$\mathbb{E}(Y^{X=x}) = \sum_{x'} \sum_{z} \mathbb{E}(Y \mid X = x', Z = z) \times \mathbb{P}(X = x') \times \mathbb{P}(Z = z \mid X = x).$$

Indeed, under the causal model of Figure A.11 (b), $Y^{X=x} \perp\!\!\!\perp X \mid W$, $Y \perp\!\!\!\perp X \mid \{W, Z\}$, $Z \perp\!\!\!\perp W \mid X$, $Y^{Z=z} \perp\!\!\!\perp Z \mid W$ and $Y^{Z=z} \perp\!\!\!\perp Z \mid X$. Thus

$$\begin{aligned}
\mathbb{E}(Y^{X=x}) &= \sum_{w} \mathbb{E}(Y \mid W = w, X = x)\mathbb{P}(W = w), \\
&= \sum_{w} \sum_{z} \mathbb{E}(Y \mid W = w, X = x, Z = z)\mathbb{P}(W = w)\mathbb{P}(Z = z \mid W = w, X = x), \\
&= \sum_{w} \sum_{z} \mathbb{E}(Y \mid W = w, Z = z)\mathbb{P}(W = w)\mathbb{P}(Z = z \mid X = x), \\
&= \sum_{z} \mathbb{E}(Y^{Z=z})\mathbb{P}(Z = z \mid X = x), \\
&= \sum_{x'} \sum_{z} \mathbb{E}(Y \mid X = x', Z = z)\mathbb{P}(X = x')\mathbb{P}(Z = z \mid X = x).
\end{aligned}$$



$$\begin{cases} X &= f_X(W, U_X), \\ M &= f_M(X, W, U_M), \\ W &= f_W(U_W), \\ Y &= f_Y(X, M, W, U_Y). \end{cases}$$

Figure A.12: Example of SCM where X is a potential cause of M and Y, where M is also a potential cause of Y, and where they are all potentially caused by W. For readability, exogenous variables U_X, U_M, U_W and U_Y have been dropped in the graphical representation.

Note that the conditions proposed through the back-door and front-door criteria are sufficient conditions for the identifiability of total causal effects. A generalization has also been proposed through several necessary and sufficient graphical conditions (Shpitser and Pearl, 2006, Tian and Pearl, 2002), along with a sound and complete identifiability algorithm (Huang and Valtorta, 2006, Shpitser and Pearl, 2006).

A.2 Mediation Analysis

Intermediary factors are sometimes involved in a given exposure-outcome relationship; consider for example the configuration depicted by the DAG given in Figure A.8 (a), where M is potentially caused by the exposure X and is also a potential cause of the outcome Y. In Section A.1 we mentioned that in that case, M is called a mediator for the $X - Y$ relationship. Moreover, if one were to adjust for the mediator M, the quantity estimated in practice would usually not be the total causal effect of X on Y, because some of the mechanisms through which X affects Y would be blocked. However, under such configuration, the total causal effect of X on Y is not the only causal effect of interest: in particular, the portion of the total causal effect of X on Y which passes indirectly through M, or in other words, which is mediated through M, can be of interest too, notably to gain further insight into the mechanisms underlying the relationship between X and Y.

A.2.1 Natural effects

A decomposition of total causal effect through natural direct and indirect effects

Let X denote the exposure variable, Y the outcome and M the possible mediator variable, which are all variables observed in the real world. As mentioned in Section A.1, a formal definition of the total causal effect of X on Y can be given based on counterfactual outcomes, notably $Y^{X=x}$ and $Y^{X=x^*}$, for any two given possible values x and x^* of X. On the other hand, in the context of mediation, the definition of the causal effects of interest will rely on more complex counterfactual variables, such as $Y^{X=x,M=m}$ and $Y^{X=x,M=M^{X=x^*}}$, for $x \neq x^*$ (Pearl, 2001, Robins and Greenland, 1992), which will be introduced below.

(a) (b)

Figure A.13: (a) Graphical representation of the causal system defined in Figure A.12 in the counterfactual world following $do(X = 0)$. (b) Same causal system in the counterfactual world following $do(X = 1)$.

Recall that $M^{X=x}$ denotes the mediator variable that would have been observed in the counterfactual world following the intervention where X would have been set to value x. Then, $Y^{X=x,M=m}$ denotes the outcome variable that would have been observed in the counterfactual world following the intervention where X and M would have been set to values x and m, respectively (for any given possible value m of M). Finally, $Y^{X=x,M=M^{x^*}}$ is the outcome variable that would have been observed in the counterfactual world following the intervention where X would have been set to value x, and where M would have been set to the value the mediator variable would have taken in the counterfactual world where X would have been set to value x^*. In particular, under a causal system such as the one whose structural causal model is given in Figure A.12, we have

- $M^{X=x^*} = f_M(x^*, W, U_M)$,
- $Y^{X=x,M=m} = f_Y(x, m, W, U_Y)$,
- $Y^{X=x,M=M^{x^*}} = f_Y(x, M^{X=x^*}, W, U_Y) = f_Y(x, f_M(x^*, W, U_M), W, U_Y)$.

Note that $Y^{X=x,M=M^{X=x}} = f_Y(x, M^{X=x}, W, U_Y) = Y^{X=x}$. Moreover, it is instructive to have a look at the representation of the causal system of Figure A.12 in counterfactual worlds. For instance, in the case where X is a binary variable, Figure A.13 (a) and Figure A.13 (b) are the systems corresponding to Figure A.12 that would be observed in the two counterfactual worlds following interventions $do(X = 1)$ and $do(X = 1)$, respectively. It is noteworthy that the counterfactual variables $Y^{X=x,M=M^{x^*}}$, for $x \neq x^*$, are not be observed in any of these two counterfactual worlds: they are said to be cross-world quantities, as they involve variables living in different counterfactual worlds.

Interestingly, the total causal effect of X on Y can be decomposed as

$$\begin{aligned}
ATE &= \mathbb{E}\left(Y^{X=1} - Y^{X=0}\right), \\
&= \mathbb{E}\left(Y^{X=1,M=M^{X=1}} - Y^{X=0,M=M^{X=0}}\right), \\
&= \mathbb{E}\left(Y^{X=1,M=M^{X=1}} - Y^{X=0,M=M^{X=1}}\right) + \mathbb{E}\left(Y^{X=0,M=M^{X=1}} - Y^{X=0,M=M^{X=0}}\right), \\
&:= NDE(1) + NIE(0),
\end{aligned}$$

or also as

$$\begin{aligned} ATE &= \mathbb{E}\left(Y^{X=1,M=M^{X=1}} - Y^{X=1,M=M^{X=0}}\right) + \mathbb{E}\left(Y^{X=1,M=M^{X=0}} - Y^{X=0,M=M^{X=0}}\right), \\ &:= NIE(1) + NDE(0), \end{aligned}$$

where $NDE(1)$ and $NDE(0)$ are two versions of the so-called "natural direct effect", while $NIE(0)$ and $NIE(1)$ are the "corresponding" versions of the so-called "natural indirect effect" (Pearl, 2001, Robins and Greenland, 1992).

Identifiability conditions for natural direct and indirect effects

Given these formal definitions of the natural direct and indirect effects, a natural question is that of their estimation from observational data. Just as for total causal effects, natural direct and indirect effects have first to be expressed as functions of observable quantities. For this purpose, sets of sufficient conditions have been established, ensuring that NIE and NDE can be identified.

First, as natural direct and indirect effects involve different counterfactual variables than total causal effects, the consistency conditions presented in Section A.1 have to be extended. In the mediation analysis setting, the consistency condition is now

(C1) If $X = x$, then $M^x = M$.

(C2) If $M^{x^*} = m$, then $Y^{x,M^{x^*}} = Y^{x,m}$.

(C3) If $X = x$ and $M = m$, then $Y^{x,m} = Y$.

Similarly, the conditional ignorability condition has to be extended. In the mediation analysis setting, it is made of the following conditions (Pearl, 2001)

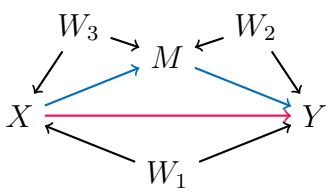
(1) $Y^{X=x,M=m} \perp\!\!\!\perp X \mid W$, for any possible values x and m of X and M, respectively.

(2) $Y^{X=x,M=m} \perp\!\!\!\perp M \mid \{X, W\}$, for any possible values x and m of X and M, respectively.

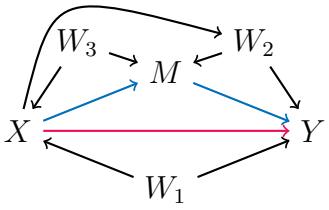
(3) $M^{X=x} \perp\!\!\!\perp X \mid W$, for any possible value x of X.

(4) $Y^{X=x,M=m} \perp\!\!\!\perp M^{X=x^*} \mid W$, for any possible values x and x^* of X and any possible value m of M.

These four conditions are usually interpreted as (1) the absence of unmeasured confounder for the $M - Y$ relationship, (2) the absence of unmeasured confounder for the $M-Y$ relationship, (3) the absence of unmeasured confounder for the $X-M$ relationship, and (4) the absence of confounder for the $M-Y$ relationship that are caused by X. These conditions are fulfilled in the example of Figure A.12, and of Figure A.14 (a) after setting $W = (W_1, W_2, W_3)$. On the other hand, they are not fulfilled in the example of Figure A.14 (b) since W_2 is a confounder of the $M-Y$ relationship, and is affected by X. Again,



(a)



(b)

Figure A.14: (a) Example of graphical causal model where the effect of X on Y is mediated through M, and where W_1 is a confounder for the $X-Y$ relationship, W_3 is a confounder for the $X-M$ relationship and W_2 is a confounder for the $M-Y$ relationship. (b) Example of graphical causal model where the effect of X on Y is mediated through M and W_2, and where W_1 is a confounder for the $X-Y$ relationship, W_3 is a confounder for the $X-M$ relationship and W_2 is a confounder for the $M-Y$ relationship.

twin networks (Balke and Pearl, 1994, Pearl, 2000) can be used to help "visualize" whether these 4 conditions, especially condition (4), hold or not under a given causal model; see below for more details. Under these extended consistency and conditional ignorability conditions, we have for any given possible values x, x^* of X,

$$\mathbb{E}(Y^{X=x,M=M^{X=x^*}}) = \sum_w \mathbb{E}(Y^{X=x,M=M^{X=x^*}} \mid W=w) \times \mathbb{P}(W=w),$$

with

$$\begin{aligned}
\mathbb{E}(Y^{X=x,M=M^{X=x^*}} \mid W=w) &= \sum_m \mathbb{E}(Y^{X=x,M=M^{X=x^*}} \mid W=w, M^{x^*}=m) \\
&\quad \times \mathbb{P}(M^{x^*}=m \mid W=w), \\
&\overset{\text{(C2)}}{=} \sum_m \mathbb{E}(Y^{X=x,M=m} \mid W=w, M^{X=x^*}=m) \\
&\quad \times \mathbb{P}(M^{x^*}=m \mid W=w), \\
&\overset{(4)}{=} \sum_m \mathbb{E}(Y^{x,m} \mid W=w) \times \mathbb{P}(M^{x^*}=m \mid W=w), \\
&\overset{(1)}{=} \sum_m \mathbb{E}(Y^{x,m} \mid W=w, X=x) \times \mathbb{P}(M^{x^*}=m \mid W=w), \\
&\overset{(2)}{=} \sum_m \mathbb{E}(Y^{x,m} \mid W=w, X=x, M=m) \\
&\quad \times \mathbb{P}(M^{x^*}=m \mid W=w), \\
&\overset{(3)}{=} \sum_m \mathbb{E}(Y^{x,m} \mid W=w, X=x, M=m) \\
&\quad \times \mathbb{P}(M^{x^*}=m \mid W=w, X=x^*), \\
&\overset{\text{(C1)}}{=} \sum_m \mathbb{E}(Y^{x,m} \mid W=w, X=x, M=m) \\
&\quad \times \mathbb{P}(M=m \mid W=w, X=x^*), \\
&\overset{\text{(C3)}}{=} \sum_m \mathbb{E}(Y \mid W=w, X=x, M=m) \\
&\quad \times \mathbb{P}(M=m \mid W=w, X=x^*).
\end{aligned}$$

De facto:

$$\begin{aligned} NDE(1) &= \mathbb{E}\left(Y^{X=1,M=M^{X=1}} - Y^{X=0,M=M^{X=1}}\right), \\ &= \sum_w \left[\mathbb{E}\left(Y^{X=1,M=M^{X=1}} \mid W=w\right) - \mathbb{E}\left(Y^{X=0,M=M^{X=1}} \mid W=w\right)\right] \times \mathbb{P}(W=w), \\ &= \sum_w \sum_m \left[\mathbb{E}\left(Y \mid W=w, X=1, M=m\right) - \mathbb{E}\left(Y \mid W=w, X=0, M=m\right)\right] \\ &\quad \times \mathbb{P}(M=m \mid W=w, X=1) \times \mathbb{P}(W=w), \end{aligned}$$

and

$$\begin{aligned} NIE(0) &= \mathbb{E}\left(Y^{X=0,M=M^{X=1}} - Y^{X=0,M=M^{X=0}}\right), \\ &= \sum_w \left[\mathbb{E}\left(Y^{X=0,M=M^{X=1}} \mid W=w\right) - \mathbb{E}\left(Y^{X=0,M=M^{X=0}} \mid W=w\right)\right] \times \mathbb{P}(W=w), \\ &= \sum_w \sum_m \mathbb{E}\left(Y \mid W=w, X=0, M=m\right) \\ &\quad \times \left[\mathbb{P}(M=m \mid W=w, X=1) - \mathbb{P}(M=m \mid W=w, X=0)\right] \\ &\quad \times \mathbb{P}(W=w). \end{aligned}$$

And in the same way

$$\begin{aligned} NDE(0) &= \sum_w \sum_m \left[\mathbb{E}\left(Y \mid W=w, X=1, M=m\right) - \mathbb{E}\left(Y \mid W=w, X=0, M=m\right)\right] \\ &\quad \times \mathbb{P}(M=m \mid W=w, X=0) \times \mathbb{P}(W=w), \end{aligned}$$

and

$$\begin{aligned} NIE(1) &= \sum_w \sum_m \mathbb{E}\left(Y \mid W=w, X=1, M=m\right) \\ &\quad \times \left[\mathbb{P}(M=m \mid W=w, X=1) - \mathbb{P}(M=m \mid W=w, X=0)\right] \\ &\quad \times \mathbb{P}(W=w). \end{aligned}$$

These quantities are expressed in terms of the distributions of the observed variables X, M, W and Y only, and can therefore be estimated in practice (provided that some positivity conditions also hold).

Twin networks may help to verify if the independence relationships given in conditions (1), (2), (3) and (4) hold under a given causal system. Consider for instance the configuration of Figure A.14 (a), where variables W_1, W_2 and W_3 are the same in the real world and counterfactual worlds following $do(X = x^*)$ or $do(X = x, M = m)$. Then by looking at Figure A.15 (a), one can see that every path between X and $Y^{X=x,M=m}$ is blocked by $\{W\} := \{W_1, W_2, W_3\}$. Moreover, every path between M and $Y^{X=x,M=m}$ is blocked by $\{W, X\}$, and conditions (1) and (2) then hold. Then in Figure A.15 (b), it is possible to

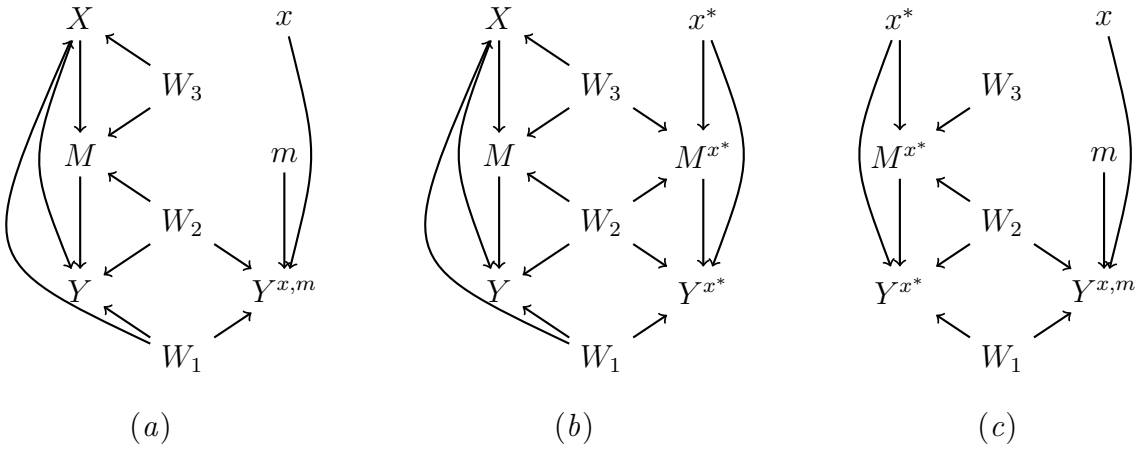


Figure A.15: (a) Twin network representation of the causal model given in Figure A.14 (a), in the real world and in the counterfactual world following the hypothetical intervention where X and M would have been set to values x and m, respectively. (b) Twin network representation of the causal model given in Figure A.14 (a), in the real world and in the counterfactual world following the hypothetical intervention where X would have been set to value x. (c) Twin network representation of the causal model given in Figure A.14 (a), in the counterfactual world following the hypothetical intervention where X would have been set to value x^*, and in the counterfactual world following the hypothetical intervention where X and M would have been set to values x and m, respectively.

see that W also blocks every path between $M^{X=x}$ and X, so that condition (3) holds. Finally, one can see in Figure A.15 (c) that W d-separates $Y^{X=x,M=m}$ and $M^{X=x^*}$, so condition (4) also holds.

On the other hand, under the configuration of Figure A.14 (b), W_2 is a confounder for the $M - Y$ relationship, and is affected by the exposure X. The fact that condition (4) does not hold under this configuration is particularly visible when looking at the twin network representation of the causal model in the counterfactual world following $do(X = x^*)$ and in the counterfactual world following $do(X = x, M = m)$, given in Figure A.16. Variables W_1 and W_3 are the same in the real world and in any counterfactual world, but this is not the case for variable W_2, which is affected by X. In the counterfactual world following $do(X = x^*)$ it has to be labeled $W_2^{X=x^*}$, and in the counterfacutal world following $do(X = x, M = m)$ it has to be labeled $W_2^{X=x}$. Note that here, exogenous variables have to be dealt with very carefully: the exogenous variable which is a direct cause of W_2 in the real world is as well a direct cause of $W_2^{X=x^*}$ and $W_2^{X=x}$ in the counterfactual worlds. This variable, denoted by U_{W_2}, is thus shared between these worlds, and here creates a path between $Y^{X=x,M=m}$ and $M^{X=x^*}$. By looking at Figure A.16, one can then see that the path $(Y^{X=x,M=m} \leftarrow W_2^x \leftarrow U_{W_2} \rightarrow W_2^{X=x^*} \rightarrow M^{X=x^*})$ in not blocked, and that $\{W\} := \{W_1, W_2, W_3\}$ cannot block it; then $Y^{X=x,M=m} \not\perp\!\!\!\perp M^{X=x^*} \mid W$. Finally note that condition (4), though maybe less evident than conditions (1), (2) or (3), actually emphasizes that W_2 should be adjusted for as it is confounding the $M - Y$ relationship, but as it is also a mediator for the $X - Y$ relationship, adjusting for W_2 would block some

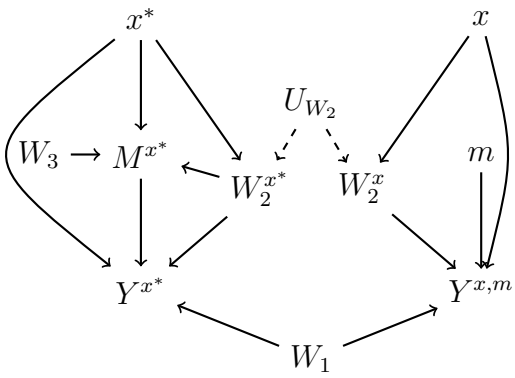


Figure A.16: Twin network representation of the causal model given in Figure A.14 (b), in the counterfactual world following the hypothetical intervention where X would have been set to value x^* and in the counterfactual world following the hypothetical intervention where X and M would have been set to values x and m, respectively. Most of the exogenous variables have been dropped for readability.

of the effect of X on Y. As a result here, and even if W_2 is observed, natural direct and indirect effects are usually not identifiable, and in that case it is not possible to determine the portion of total effect of X on Y which passes through M (Avin et al., 2005).

Several extensions have been introduced in the literature to deal with multiple mediators, notably when certain mediators are also confounders for the relationship between other mediators and the outcome. A first solution is to gather all the mediators in one multivariate mediator variable, and then to study the effect of the exposure on the outcome through this unique mediator. However, this method does not allow to assess the portion of the total causal effect which is mediated through a specific mediator. Alternatively, when conditions $(1) - (2)$ and (3) hold for the mediator of interest, but not condition (4), a solution is to consider randomized interventional analogues of the natural direct and indirect effects (VanderWeele, 2015, VanderWeele et al., 2014).

A.3 Longitudinal Models

A.3.1 Preamble

So far, the causal models that we considered involved time-fixed variables only, even if implicitly, for instance in the configuration given in Figure A.12, W was supposed to be anterior to X, M and Y, X was supposed to be anterior to M and Y, and M anterior to Y. However, for certain exposures, and in particular for lifestyle exposures such as obesity, the true causal model is likely to involve time-varying variables.

Consider for instance the causal relationship between obesity and cancer occurrence. Because insufficient physical activity is likely to increase the risk of obesity, as well as

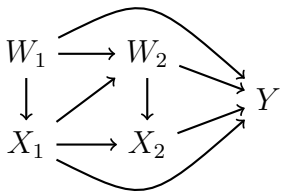


Figure A.17: Example of (graphical) longitudinal causal model, with time-varying exposure $(X_t)_{t\in[\![1;2]\!]}$ and time-varying confounder $(W_t)_{t\in[\![1;2]\!]}$ affected by the exposure.

to increase the risk of cancer, physical activity is usually considered as a confounder in the obesity-cancer relationship. But on the other hand, obesity is also likely to decrease physical activity, and then physical activity could be considered as mediator in the obesity-cancer relationship. Then, the causal models given in Figure A.8 (a) or Figure A.9 (a) are too simplistic to properly describe the relationship between obesity and physical activity. The true causal model actually involves time-varying variables, such as the one depicted in Figure A.17. A causal model with time-varying exposure, mediators and confounders is called a longitudinal causal model; usually, the setting of discrete times is considered. In a longitudinal causal model, the value of the exposure variable at a time t may affect the future value of the confounding variable, at a any time $t_1 > t$, as well as its own future value; then the distinction between mediators and confounders is not as clear as before. For instance in the configuration of Figure A.17, W_1 is a confounder for the $X_1 - Y$ and $X_2 - Y$ relationships, W_2 is a confounder for the $X_2 - Y$ relationship, but also a mediator for the $X_1 - Y$ relationship.

The tools and principles of causal inference presented in Section A.1 or in Section A.2 for mediation analysis have been extended to longitudinal settings. They are presented in Section A.3.3 for total causal effects, while we refer to VanderWeele (2015) for the extensions of natural direct and indirect effects.

A.3.2 Notations in longitudinal settings

We will here work under the standard setting where the time-varying variables are observable at discrete times over a time-window $[\![1;T]\!] := \{1,\dots,T\}$, for some time $T > 1$, and where the outcome Y is measured at time point T. For any time $t \in [\![1;T]\!]$, we let X_t denote the exposure variable at time t, and then adopt the following notation (Daniel et al., 2012, Hernán and Robins, 2020, VanderWeele, 2015): $\bar{X}_t = (X_1, X_2, \dots, X_t)$ denotes the exposure profile until time t, $\bar{X} = \bar{X}_T = (X_1, X_2, \dots X_T)$ denotes the "full" exposure profile, while $\bar{x}_t$ or $\bar{x}$ denote specific (fixed) profiles for the exposure up to time t or T, respectively. We further let $\bar{X}_t$ denote the empty set when $t < 1$. When needed, we will use similar notations for mediator processes $(M_t)_{t\geq 1}$, as well as confounder processes $(W_t)_{t\geq 1}$. Finally, to simplify we assume that all the variables are binary.

A.3.3 Total causal effects

Extension of the definition of total causal effects

The formal definition of total causal effect for time-fixed exposures is extended to time-varying exposures by comparing counterfactual outcomes related to hypothetical interventions on the exposure over a time-interval. Let $Y^{\bar{X}=\bar{x}}$ denote the outcome variable (at time T) that would have been observed in the counterfactual world where $\bar{X}$ would have been set to a given possible profile $\bar{x}$. Then $\mathbb{E}\left(Y^{\bar{X}=\bar{x}} - Y^{\bar{X}=\bar{x}^*}\right)$, for any given profiles $\bar{x}$ and $\bar{x}^*$, is one measure of the total causal effect of the exposure until time T on the outcome Y. Several comparisons of profiles are possibles, depending on the choice of $\bar{x}$ and $\bar{x}^*$; for instance if the exposure of interest is the obesity status, one could compare the profiles $(1, \ldots, 1)$ and $(0, \ldots, 0)$, namely the profiles "always obese" and "never obese". On the other hand, the quantity $\mathbb{E}\left(Y^{X_t=x_t} - Y^{X_t=x_t^*}\right)$, for any time $t \in [\![1; T]\!]$, also has a clear causal meaning: it is the total causal effect of the exposure at time t on the outcome Y. However, quantities like $\mathbb{E}\left(Y^{\bar{X}=\bar{x}} - Y^{\bar{X}=\bar{x}^*}\right)$ are usually preferred when working with time-varying exposures, because they account for the variations of the exposure variable over time.

Identifiability conditions for total causal effects of a time-varying exposure

Once again, to estimate such a counterfactual quantity from observational data, it has first to be expressed as a function of the observations. For this purpose, the sets of sufficient identifiablity conditions proposed for time-fixed variables (Robins, 1986, Rosenbaum and Rubin, 1983) and presented in Section A.1 have been extended.

First to connect conterfactual variables such as $Y^{\bar{X}=\bar{x}}$ to observed variables, the consistency condition is now (Daniel et al., 2012)

$$\text{(C*)} \qquad \text{if } \bar{X} = \bar{x}, \text{ then } Y^{\bar{X}=\bar{x}} = Y.$$

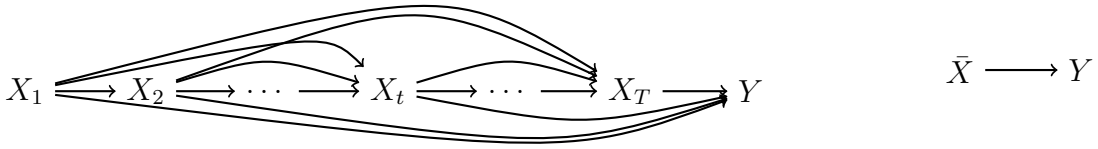
Then, the first set of sufficient conditions includes the unconditional ignorability condition, formulated relatively to the full exposure profile $\bar{X}$

$$Y^{\bar{X}=\bar{x}} \perp\!\!\!\perp \bar{X}, \text{ for any possible profile } \bar{x} \text{ of } \bar{X}.$$

Similarly, a second set of identifiability conditions may be formulated with a conditional version of the ignorablity condition

$$Y^{\bar{X}=\bar{x}} \perp\!\!\!\perp \bar{X} \mid \bar{W}, \text{ for any possible profile } \bar{x} \text{ of } \bar{X}.$$

Note that these two sets of sufficient conditions are simply a reformulation of the conditions presented in Section A.1, when the exposure of interest if $\bar{X}$. However, they



(a) (b)

Figure A.18: (a) Longitudinal causal model with a time-varying exposure $(X_t)_{t\in[\![1;T]\!]}$, in the absence of confounding. (b) Compact representation of the longitudinal causal model given in Figure A.18 (a).

hold only under very simplistic longitudinal configurations. For example consider the simple longitudinal configuration given in Figure A.18 (a), where the time-varying exposure $(X_t)_{t\in[\![1;T]\!]}$ is a potential cause of outcome Y; for any time $t \in [\![1; T-1]\!]$, X_t may further affect X_{t_1}, for $t_1 > t$, but no variable is confounding the $\bar{X} - Y$ relationship. A more compact representation of the model is also given in Figure A.18 (b). Then the unconditional ignorability condition holds, and for any given $\bar{x}$ and $\bar{x}^*$ in $\{0,1\}^T$

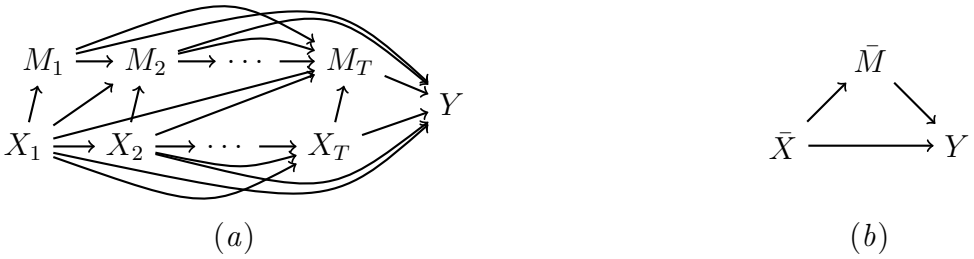
$$\begin{aligned} ATE\left(\bar{x};\bar{x}^*\right) &:= \mathbb{E}\left(Y^{\bar{X}=\bar{x}}\right) - \mathbb{E}\left(Y^{\bar{X}=\bar{x}^*}\right), \\ &= \mathbb{E}\left(Y^{\bar{X}=\bar{x}} \mid \bar{X}=\bar{x}\right) - \mathbb{E}\left(Y^{\bar{X}=\bar{x}^*} \mid \bar{X}=\bar{x}^*\right), \\ &\overset{(C^*)}{=} \mathbb{E}\left(Y \mid \bar{X}=\bar{x}\right) - \mathbb{E}\left(Y \mid \bar{X}=\bar{x}^*\right). \end{aligned}$$

On the other hand, the unconditional ignorability condition also holds in the presence of a time-varying "pure" mediator, as in Figure A.19 (a). Then, assume the presence of a time-varying "pure" confounder, as in Figure A.20 (a); $\bar{W}$ is confounding the $\bar{X} - Y$ relationship and thus for any given $\bar{x}$ and $\bar{x}^*$ in $\{0,1\}^T$

$$\begin{aligned} ATE\left(\bar{x};\bar{x}^*\right) &= \sum_{\bar{w}} \left[\mathbb{E}\left(Y^{\bar{X}=\bar{x}} \mid \bar{W}=\bar{w}\right) - \mathbb{E}\left(Y^{\bar{X}=\bar{x}^*} \mid \bar{W}=\bar{w}\right)\right] \times \mathbb{P}(\bar{W}=\bar{w}), \\ &= \sum_{\bar{w}} \left[\mathbb{E}\left(Y^{\bar{X}=\bar{x}} \mid \bar{W}=\bar{w}, \bar{X}=\bar{x}\right) - \mathbb{E}\left(Y^{\bar{X}=\bar{x}^*} \mid \bar{W}=\bar{w}, \bar{X}=\bar{x}^*\right)\right] \\ &\qquad \times \mathbb{P}(\bar{W}=\bar{w}), \\ &\overset{(C^*)}{=} \sum_{\bar{w}} \left[\mathbb{E}\left(Y \mid \bar{W}=\bar{w}, \bar{X}=\bar{x}\right) - \mathbb{E}\left(Y \mid \bar{W}=\bar{w}, \bar{X}=\bar{x}^*\right)\right] \times \mathbb{P}(\bar{W}=\bar{w}), \end{aligned}$$

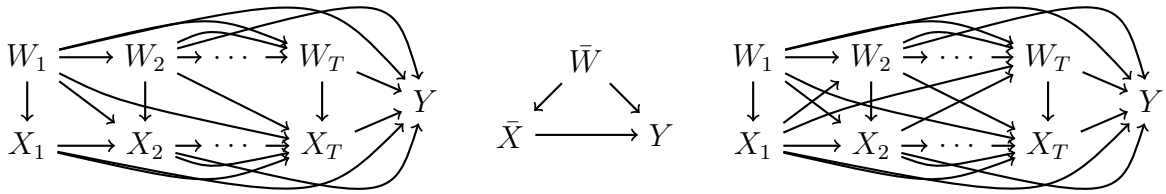
where the sum is over all possible values of $\bar{W}$.

It is noteworthy that neither the unconditional ignorability condition nor the conditional ignorability condition holds under the causal model of Figure A.20 (c). However, a "sequential" version of the ignorability condition (Robins, 1986) holds under this configuration, so that the total causal effect of the time-varying exposure can still be identified.



(a) (b)

Figure A.19: (a) Longitudinal causal model with a time-varying exposure $(X_t)_{t\in[\![1;T]\!]}$ and time-varying "pure" mediator $(M_t)_{t\in[\![1;T]\!]}$, in the absence of confounding. (b) Compact representation of the longitudinal causal model given in Figure A.19 (a).



(a) (b) (c)

Figure A.20: (a) Longitudinal causal model with a time-varying exposure $(X_t)_{t\in[\![1;T]\!]}$ and time-varying "pure" confounder $(W_t)_{t\in[\![1;T]\!]}$. (b) Compact representation of the longitudinal causal model given in Figure A.20 (a). (c) Longitudinal causal model with a time-varying exposure $(X_t)_{t\in[\![1;T]\!]}$ and time-varying confounder $(W_t)_{t\in[\![1;T]\!]}$ affected by the exposure.

Indeed, under the causal model of Figure A.20 (c),

$$Y^{\bar{X}=\bar{x}} \perp\!\!\!\perp X_1 \mid W_1 \text{ and}$$
$$Y^{\bar{X}=\bar{x}} \perp\!\!\!\perp X_t \mid \left\{\bar{X}_{t-1}, \bar{W}_t\right\}, \text{ for any time } t \in [\![2, T]\!] \text{ and any possible profile } \bar{x} \text{ of } \bar{X}.$$

Then, for any given $\bar{x}$ in $\{0,1\}^T$ (Robins, 1986)

$$\mathbb{E}(Y^{\bar{X}=\bar{x}}) = \sum_{\bar{w}} \left[\mathbb{E}(Y \mid \bar{W}=\bar{w}, \bar{X}=\bar{x}) \times \prod_{t=0}^{T} \mathbb{P}(W_t = w_t \mid \bar{W}_{t-1}=\bar{w}_{t-1}, \bar{X}_{t-1}=\bar{x}_{t-1}) \right],$$

where the sum is over all possible values of $\bar{W}$. This expression is usually called the "G-computation algorithm formula" or "g-computation formula" (Daniel et al., 2012, Robins, 1986), and can be proven by using mathematical induction (along with the consistency condition (C*)). Under the causal model of Figure A.20 (c), $ATE\left(\bar{x};\bar{x}^*\right)$, for any $\bar{x}$ and $\bar{x}^*$ in $\{0,1\}^T$, can then be identified as

$$ATE\left(\bar{x};\bar{x}^*\right) = \sum_{\bar{w}} \Big[\mathbb{E}\left(Y \mid \bar{W}=\bar{w}, \bar{X}=\bar{x}\right) \times \prod_{t=1}^{T} \mathbb{P}\left(W_t = w_t \mid \bar{W}_{t-1}=\bar{w}_{t-1}, \bar{X}_{t-1}=\bar{x}_{t-1}\right)$$
$$- \mathbb{E}\left(Y \mid \bar{W}=\bar{w}, \bar{X}=\bar{x}^*\right) \times \prod_{t=1}^{T} \mathbb{P}\left(W_t = w_t \mid \bar{W}_{t-1}=\bar{w}_{t-1}, \bar{X}_{t-1}=\bar{x}^*_{t-1}\right) \Big].$$

Appendix B

Preliminary results on mediation analysis under over-simplified longitudinal models

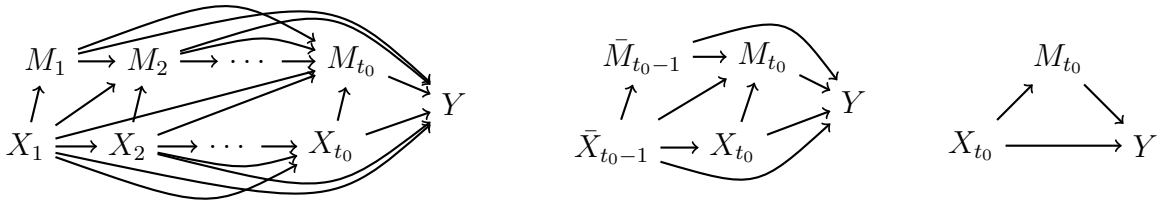
The aim of this Appendix is to present preliminary results on mediation analysis, regarding the problematic presented in Chapter 3 for total causal effects. We here turn our attention to the decomposition of the total causal effect of the exposure of interest into the sum of (i) an indirect effect through given potential mediator(s), and (ii) a direct effect. We will focus on the decomposition based on the so-called natural direct and indirect effects (Pearl, 2001, Robins and Greenland, 1992, VanderWeele, 2015). We consider simple longitudinal causal models and study whether mediation analysis performed under over-simplification of these causal models may produce valid results. We start with the situation where only levels of exposures measured at recruitment are available, and we will then briefly present some results in the situation where summary measures of past levels of exposures are available. For the sake of conciseness, most technical details are not provided; they follow arguments and techniques similar to those used in the context of total causal effects in Chapter 3.

B.1 When instantaneous levels of exposures are available

B.1.1 Natural Effects in the absence of confounding

In this Section, we consider a simple longitudinal causal model involving a pure mediator process, as depicted in Figure B.1 (*L.Med*) or in Figure B.1 (*L.Med.compact*) in a more compact form.

For time-varying exposures and mediators, natural direct and indirect effects are de-



(*L.Med*) (*L.Med.compact*) (*CS.Med*)

Figure B.1: (*L.Med*) Longitudinal model with time-varying exposure, time-varying pure mediator, and no confounder. (*L.Med*) Compact representation of model given in Figure B.1 (*L.Med*). (*CS.Med*) Over-simplified cross-sectional model associated with the longitudinal model given in Figure B.1 (*L.Med*).

fined by

$$\begin{aligned} NDE_{L.Med}(\bar{x}_{t_0};\bar{x}^*_{t_0}) &= \mathbb{E}_{L.Med}\Big(Y^{\bar{X}_{t_0}=\bar{x}_{t_0},\bar{M}_{t_0}=\bar{M}^{\bar{x}_{t_0}}_{t_0}} - Y^{\bar{X}_{t_0}=\bar{x}^*_{t_0},\bar{M}_{t_0}=\bar{M}^{\bar{x}_{t_0}}_{t_0}}\Big), \\ NIE_{L.Med}(\bar{x}_{t_0};\bar{x}^*_{t_0}) &= \mathbb{E}_{L.Med}\Big(Y^{\bar{X}_{t_0}=\bar{x}^*_{t_0},\bar{M}_{t_0}=\bar{M}^{\bar{x}_{t_0}}_{t_0}} - Y^{\bar{X}_{t_0}=\bar{x}^*_{t_0},\bar{M}_{t_0}=\bar{M}^{\bar{x}^*_{t_0}}_{t_0}}\Big), \end{aligned}$$

for two given profiles $\bar{x}_{t_0}$ and $\bar{x}^*_{t_0}$ in $\{0,1\}^{t_0}$ of the exposure, and with, for example, $\bar{M}^{\bar{x}_{t_0}}_{t_0}$ denoting the mediator profile that would be observed in the counterfactual world following $do(\bar{X}_{t_0} = \bar{x}_{t_0})$. We refer to Pearl (2001), Robins and Greenland (1992), VanderWeele (2015), VanderWeele and Tchetgen Tchetgen (2017) for generalities on mediation analysis, including mediation analysis with time-varying exposures and mediators. Under the model given in Figure B.1 (*L.Med*), the natural direct and indirect effects, $NDE_{L.Med}$ and $NIE_{L.Med}$, are identified as

$$\begin{aligned} NDE_{L.Med}\left(\bar{x}_{t_0};\bar{x}^*_{t_0}\right) = \sum_{\bar{m}_{t_0}} &\left[\mathbb{E}\left(Y \mid \bar{X}_{t_0}=\bar{x}_{t_0},\bar{M}_{t_0}=\bar{m}_{t_0}\right) - \mathbb{E}\left(Y \mid \bar{X}_{t_0}=\bar{x}^*_{t_0},\bar{M}_{t_0}=\bar{m}_{t_0}\right)\right] \\ &\times \mathbb{P}\left(\bar{M}_{t_0}=\bar{m}_{t_0} \mid \bar{X}_{t_0}=\bar{x}_{t_0}\right), \end{aligned}$$

and

$$\begin{aligned} NIE_{L.Med}\left(\bar{x}_{t_0};\bar{x}^*_{t_0}\right) = \sum_{\bar{m}_{t_0}} &\mathbb{E}\left(Y \mid \bar{X}_{t_0}=\bar{x}_{t_0},\bar{M}_{t_0}=\bar{m}_{t_0}\right) \\ &\times \left[\mathbb{P}\left(\bar{M}_{t_0}=\bar{m}_{t_0} \mid \bar{X}_{t_0}=\bar{x}_{t_0}\right) - \mathbb{P}\left(\bar{M}_{t_0}=\bar{m}_{t_0} \mid \bar{X}_{t_0}=\bar{x}^*_{t_0}\right)\right], \end{aligned}$$

and can then be estimated, provided that data on $\bar{X}_{t_0}$ and $\bar{M}_{t_0}$ are available and that some positivity condition holds. But, when the exposure of interest and the mediator are only measured at time t_0, practitioners usually overlook their time-varying natures, and work under the over-simplified causal model depicted in Figure B.1 (*CS.Med*). They would then usually consider $NDE_{CS.Med} = \mathbb{E}_{CS.Med}\big(Y^{X_{t_0}=1,M_{t_0}=M^1_{t_0}} - Y^{X_{t_0}=0,M_{t_0}=M^1_{t_0}}\big)$ and $NIE_{CS.Med} = \mathbb{E}_{CS.Med}\big(Y^{X_{t_0}=0,M_{t_0}=M^1_{t_0}} - Y^{X_{t_0}=0,M_{t_0}=M^0_{t_0}}\big)$, instead of $NDE_{L.Med}$ and $NIE_{L.Med}$. Under model (*CS.Med*), $Y^{X_{t_0}=x_{t_0},M_{t_0}=m_{t_0}} \perp\!\!\!\perp \{X_{t_0}, M_{t_0}\}$, $Y^{X_{t_0}=x_{t_0},M_{t_0}=m_{t_0}} \perp\!\!\!\perp$

$M_{t_0}^{X_{t_0}=x_{t_0}^*}$ and $M_{t_0}^{X_{t_0}=x_{t_0}^*} \perp\!\!\!\perp X_{t_0}$. As a result, it is easy to show that (Pearl, 2001)

$$\begin{aligned} NDE_{CS.Med} &\fallingdotseq \sum_{m_{t_0}} [\mathbb{E}(Y \mid X_{t_0}=1, M_{t_0}=m_{t_0}) - \mathbb{E}(Y \mid X_{t_0}=0, M_{t_0}=m_{t_0})] \\ &\quad \times \mathbb{P}(M_{t_0}=m_{t_0} \mid X_{t_0}=1), \\ NIE_{CS.Med} &\fallingdotseq \sum_{m_{t_0}} \mathbb{E}(Y \mid X_{t_0}=0, M_{t_0}=m_{t_0}) \\ &\quad \times [\mathbb{P}(M_{t_0}=m_{t_0} \mid X_{t_0}=1) - \mathbb{P}(M_{t_0}=m_{t_0} \mid X_{t_0}=0)]\,. \end{aligned}$$

We recall that symbol $\fallingdotseq$ was introduced in Section 3.2 in Chapter 3.

However, model (*CS.Med*) is generally misspecified under model (*L.Med*) since $\bar{X}_{t_0-1}$ is a confounder for the $X_{t_0} - Y$, $M_{t_0} - Y$ and $X_{t_0} - M_{t_0}$ relationships, and $\bar{M}_{t_0-1}$ is a confounder for the $M_{t_0} - Y$ relationship. In particular, we can show that under model (*L.Med*), neither $NDE_{CS.Med}$ nor $NIE_{CS.Med}$ expresses as an average of longitudinal (in)direct effects. In other words, $NDE_{CS.Med}$ and $NIE_{CS.Med}$ generally have to be interpreted with caution if the true model is (*L.Med*). Turning our attention to special cases, we can show that the interpretation of $NDE_{CS.Med}$ and $NIE_{CS.Med}$ remains unclear even if both processes $(X_t)_t$ and $(M_t)_t$ are stable (with stability defined as in Section 3.3 in Chapter 3). Moreover, under the complete mediation case (when the effect of the exposure process on the outcome is entirely mediated by the mediator process), $NDE_{CS.Med}$ generally differs from zero and is therefore misleading. Interestingly, the case of absence of mediation is more subtle; we study this special case in more details in the following Section.

B.1.2 Natural effects in the absence of confounding - absence of mediation

Under model (*L.Med*), the absence of mediation arises in the case of (*i*) the absence of an effect of the exposure on the mediator, as depicted in Figures B.2 (*a*) and (*c*) and/or (*ii*) the absence of an effect of the mediator on the outcome, as depicted in Figures B.2 (*b*) and (*c*). Of course, $NIE_{L.Med}(\bar{x}_{t_0}; \bar{x}_{t_0}^*)$ equals zero in both cases, for any given profiles $\bar{x}_{t_0}$ and $\bar{x}_{t_0}^*$. Regarding $NIE_{CS.Med}$, it can be shown that it is null under case (*a*) (and (*c*)) but generally not under case (*b*). This is because $\bar{X}_{t_0-1}$ is not only a confounder for the $X_{t_0} - Y$ relationship, but also for the $X_{t_0} - M_{t_0}$ relationship under case (*b*), and as a result $Y \not\perp\!\!\!\perp M_{t_0} \mid X_{t_0}$. It can further be shown that the nullity of $NIE_{CS.Med}$ under case (*a*) (and (*c*)) is still guaranteed in the presence of an observed time-invariant confounder. However, the nullity of $NIE_{CS.Med}$ is not guaranteed anymore if the true causal model involves a time-varying confounder observed at inclusion only.

We now present numerical examples to illustrate the magnitude of $NIE_{CS.Med}$ in the absence of mediation under case (*b*). Here, we consider a simple longitudinal causal model with $t_0 = 2$ and with Gaussian variables, as it allows the derivation of closed form

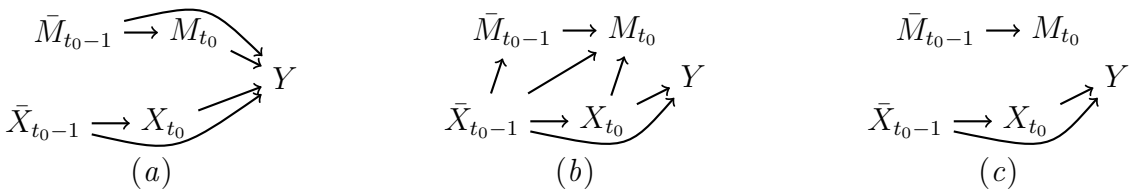


Figure B.2: Absence of mediation under model ($L.Med$) of Figure B.1: (a) Exposure process $(X_t)_t$ has no effect on $(M_t)_t$. (b) The process $(M_t)_t$ has no effect on Y. (c) Combination of the previous two cases.

expressions for NIE_{CS} (see below). More precisely, we assume that X_1, ε_{X_2}, ε_{M_1}, ε_{M_2} and ε_Y are four independent $\mathcal{N}(0,1)$ random variables, and that the structural causal model defining variables X_2, M_1, M_2 and Y is

$$\begin{aligned} X_2 &= \delta_X X_1 + \varepsilon_{X_2}, \\ M_1 &= \alpha_1 X_1 + \varepsilon_{M_1}, \\ M_2 &= \alpha_2 X_2 + \delta_M M_1 + \varepsilon_{M_2}, \\ Y &= \gamma_1 X_1 + \gamma_2 X_2 + \varepsilon_Y, \end{aligned} \tag{B.1}$$

for some δ_X, α_1, α_2, δ_M, γ_1 and γ_2 in $\mathbb{R}$. The structural equation defining the outcome Y in Equation (B.1) involves neither M_1 nor M_2 so that this causal model is an example of case (b) (with continuous X_t and M_t, for $t \in \{1,2\}$, and under the special case where $M_2 \perp\!\!\!\perp X_1 \mid \{X_2, M_1\}$). We can show that, for any $x_2 \neq x_2^*$,

$$NIE_{CS.Med}(x_2; x_2^*) \fallingdotseq \frac{\gamma_1\alpha_1\delta_M(x_2 - x_2^*)\left[\alpha_2(1+\delta_X^2) + \alpha_1\delta_M\delta_X\right]}{(1+\delta_X^2)\left[1+\delta_M^2(1+\alpha_1^2)+\delta_X^2(1+\delta_M^2)\right]}, \tag{B.2}$$

which is typically non-null. Figure B.3 illustrates the behavior of $NIE_{CS.Med}$ as a function of $\delta_X \in [\![-10, 10]\!]$, $\delta_M \in \{-5, -2, -1, 0, 1, 2, 5\}$, $\alpha_1 = \alpha_2 \in \{0, 1.25, 2.5, 3.75, 5\}$ and for the particular choices $\gamma_1 = 0.8$ and $x_2 - x_2^* = 1$. Similar results were obtained for other values of γ_1 and $x_2 - x_2^*$. Figure B.3 especially illustrates that $NIE_{CS.Med}$ is zero when X_1 is not a confounder of the $M_2 - Y$ relationship, which is the case when (i) X_1 does not cause M_2 ($\alpha_1 = \alpha_2 = 0$ or $\delta_M = 0$) or (ii) X_1 is not a direct cause of Y ($\gamma_1 = 0$). Figure B.3 also illustrates that $NIE_{CS.Med}$ is a non-monotonic function of δ_X and that $NIE_{CS.Med} \to 0$ as $|\delta_X| \to \infty$. This latter result can be explained by the fact that under case (b), when $t_0 = 2$, we should have $Y \perp\!\!\!\perp M_2 \mid \{X_1, X_2\}$, but $Y \not\perp\!\!\!\perp M_2 \mid X_2$; however, as $\mathrm{Cor}(X_1, X_2) \to 1$ when $|\delta_X| \to \infty$, Y then tends to be independent of M_2 given X_2.

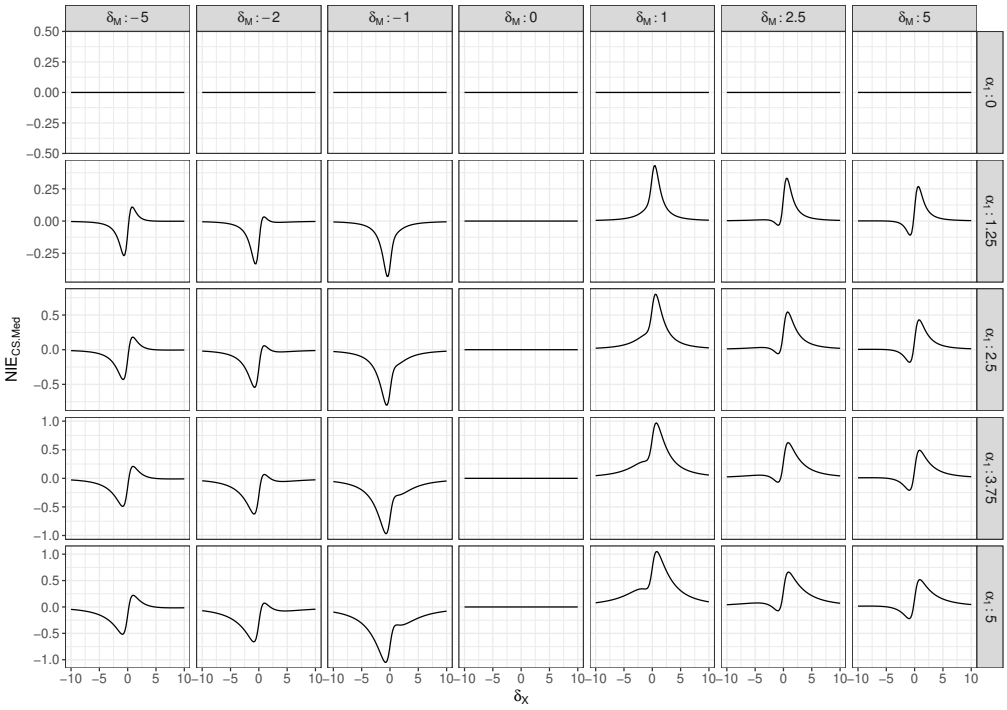


Figure B.3: Analytic values of $NIE_{CS.Med}(1;0)$, for $\delta_X \in [\![-10, 10]\!]$, $\delta_M \in \{-5, -2, -1, 0, 1, 2, 5\}$, $\alpha_1 = \alpha_2 \in \{0, 1.25, 2.5, 3.75, 5\}$, $\gamma_1 = 0.8$ and $x_2 - x_2^* = 1$, under the causal model described in Equation (B.1).

B.2 When summary variables of past levels of exposures are available

In this Section, we briefly study the causal model given in Figure B.4 (*L.Med*). Keeping in mind that versions of treatment $\mathscr{X} = \boldsymbol{x}$ are relevant under this model (see Section 3.4 in Chapter 3), the two following quantities can be considered to be of interest:

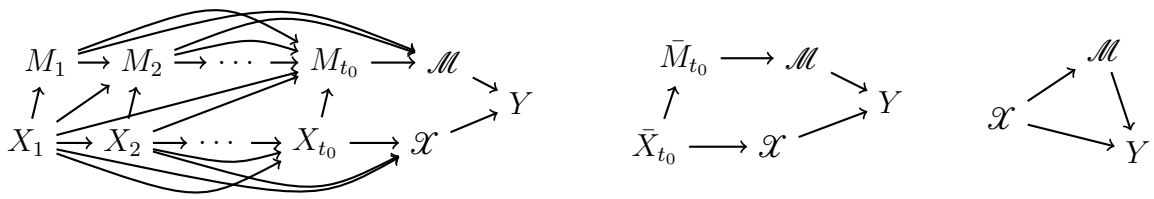
$$\sum_{\bar{x}_{t_0}} NDE_{L.Med}(\bar{x}_{t_0}; \bar{x}^*_{t_0}) \times \mathbb{P}\left(\bar{X}_{t_0} = \bar{x}_{t_0} \mid \mathscr{X} = \boldsymbol{x}\right), \tag{B.3}$$

for any $\bar{x}^*_{t_0}$ such that $\mathscr{X} = \boldsymbol{x}^*$, and

$$\sum_{\bar{x}_{t_0}} \sum_{\bar{x}^*_{t_0}} NIE_{L.Med}(\bar{x}_{t_0}; \bar{x}^*_{t_0}) \times \mathbb{P}\left(\bar{X}_{t_0} = \bar{x}_{t_0} \mid \mathscr{X} = \boldsymbol{x}\right) \times \mathbb{P}\left(\bar{X}_{t_0} = \bar{x}^*_{t_0} \mid \mathscr{X} = \boldsymbol{x}^*\right), \tag{B.4}$$

for two profiles $\bar{x}_{t_0}$ and $\bar{x}^*_{t_0}$ leading to $\mathscr{X} = \boldsymbol{x}$ and $\mathscr{X} = \boldsymbol{x}^*$, respectively.

When only data on $\mathscr{X}$ and $\mathscr{M}$ (and Y) are considered, many practitioners would work under the over-simplified causal model depicted in Figure B.4 (*SV.Med*), and would then want to estimate $NDE_{SV.Med}(\boldsymbol{x}; \boldsymbol{x}^*) = \mathbb{E}_{SV.Med}\left(Y^{\boldsymbol{x}, \mathscr{M}^{\boldsymbol{x}}} - Y^{\boldsymbol{x}^*, \mathscr{M}^{\boldsymbol{x}}}\right)$ and $NIE_{SV.Med}(\boldsymbol{x}; \boldsymbol{x}^*) = \mathbb{E}_{SV.Med}\left(Y^{\boldsymbol{x}^*, \mathscr{M}^{\boldsymbol{x}}} - Y^{\boldsymbol{x}^*, \mathscr{M}^{\boldsymbol{x}^*}}\right)$, for any $\boldsymbol{x} \neq \boldsymbol{x}^*$. If model (*SV.Med*) were true, we would



(*L.Med*) (*L.compact*) (*SV.Med*)

Figure B.4: (*L.Med*) Longitudinal model with time-varying exposure (X_t) and time-varying mediator (M_t) not affecting the exposure, that potentially affect the outcome Y only through some summary variables $\mathscr{X}$ and $\mathscr{W}$. (*L.compact*) Simplified representation of model (*L.Med*) given in Figure B.4. (*SV.Med*) Over-simplified model associated with the longitudinal model given in Figure B.4 (*L.Med*).

have $Y^{\mathscr{X}=x,\mathscr{M}=m} \perp\!\!\!\perp \{\mathscr{X}, \mathscr{M}\}$, $Y^{\mathscr{X}=x,\mathscr{M}=m} \perp\!\!\!\perp \mathscr{M}^{\mathscr{X}=x^*}$ and $\mathscr{M}^{\mathscr{X}=x^*} \perp\!\!\!\perp \mathscr{X}$. Consequently,

$$\begin{aligned} NDE_{SV.Med}(x; x^*) &\leftrightarrows \sum_{m} \left[\mathbb{E}\left(Y \mid \mathscr{X}=x, \mathscr{M}=m\right) - \mathbb{E}\left(Y \mid \mathscr{X}=x^*, \mathscr{M}=m\right)\right] \\ &\quad \times \mathbb{P}\left(\mathscr{M}=m \mid \mathscr{X}=x\right), \\ NIE_{SV.Med}(x; x^*) &\leftrightarrows \sum_{m} \mathbb{E}\left(Y \mid \mathscr{X}=x^*, \mathscr{M}=m\right) \\ &\quad \times \left[\mathbb{P}\left(\mathscr{M}=m \mid \mathscr{X}=x\right) - \mathbb{P}\left(\mathscr{M}=m \mid \mathscr{X}=x^*\right)\right], \end{aligned}$$

Even if model (*SV.Med*) is generally misspecified ($\mathscr{X}$ does not cause $\mathscr{M}$ under the model of Figure B.4 (*L.Med*) and $\left\{\bar{X}_{t_0}, \bar{M}_{t_0}\right\}$ is confounding the $\mathscr{X} - Y$ and the $\mathscr{M} - Y$ relationships), $NDE_{SV.Med}(x; x^*)$ and $NIE_{SV.Med}(x; x^*)$ actually equal the quantities given in Equations (B.3) and (B.4), respectively. However, and as already mentioned in Chapter 3 for weighted averages of longitudinal total effects, we shall stress that the interpretability of such quantity is not always straightforward. Therefore, under a configuration such as the one given in Figure B.4 (*L.Med*), considering $\mathscr{X}$ and $\mathscr{M}$ only and working under model (*SV.Med*) can be sufficient not only to estimate the total causal effect, but also to infer the amount of this effect that is mediated by $\bar{M}_{t_0}$, provided there is a certain homogeneity in the "individual" longitudinal effects. If not, the quantity estimated in practice has to be interpreted with caution. It can be shown that these results extend to the case where a time-invariant pure confounder is present: both $NDE_{SV.Med}$ and $NIE_{SV.Med}$ express as weighted averages of stratum specific natural direct and indirect effects, with strata defined according to the levels of the confounder. However, if the pure confounder is time-varying, summary measures are not sufficient anymore to recover meaningful natural direct and indirect effects; we recall that this was already the case for the total effect; see Section 3.4 in Chapter 3.

Développements méthodologiques autour de l'inférence causale et de l'analyse de données en grande dimension

Résumé : L'objectif de cette thèse est d'explorer certains enjeux soulevés par la mise en application en épidemiologie du cancer des outils développés en inférence causale. Tout d'abord, nous étudions comment l'effet d'une intervention hypothétique sur l'exposition d'intérêt, lorsque celle-ci n'est pas applicable en pratique, est lié aux effets des interventions sur certaines de ses causes. Ensuite, nous déterminons des conditions assurant que les quantités obtenues en travaillant sous de modèles causaux simplifiés, où la nature longitudinale des variables est négligée, soient liées à celles d'intérêt sous le vrai modèle longitudinal. Par ailleurs, nous étudions des modèles proposant des formulations probabilistes de techniques de réduction de dimension classiques, et identifions un défaut rencontré dans plusieurs de ces modèles. Nous nous intéressons en particulier à la formulation probabiliste des moindres carrés partiels proposée par el Bouhaddani et al. (2018) : nous décrivons en détail le défaut sous leur modèle, et l'illustrons au moyen de simulations. Enfin nous nous intéressons à la sélection du paramètre de régularisation dans le cas du lasso adaptatif. Nous montrons de manière empirique que la validation croisée «K-fold», bien que couramment employée, n'est pas adaptée à la calibration du paramètre de régularisation pour le lasso adaptatif. Nous proposons une procédure alternative, puis montrons sur des simulations qu'elle présente de meilleures performances que la validation croisée «K-fold».

Mots clés : Inférence causale, analyse en médiation, analyse de données en grande dimension.

Methodological developments around causal inference and the analysis of high-dimensional data

Abstract: The objective of this thesis is to explore some of the problematics raised by the practical application of causal inference in cancer epidemiology. First, we show how the effect of an hypothetical intervention on the exposure of interest, when impossible to apply in practice, relates to the effects of interventions on its causes, depending on the structure of the causal model. Second, we investigate conditions ensuring that estimates derived under over-simplified causal models, where the longitudinal nature of the variables have been neglected, relate to causal quantities of interest under the true longitudinal causal model. Then, we study several models proposing probabilistic formulations of dimension-reduction techniques, where we identify a defect. We focus in particular on the probabilistic formulation of partial least squares proposed by el Bouhaddani et al. (2018): we describe the limitation under their model and further illustrate it through simulated examples. Finally, we study the calibration of the tuning parameter in the adaptive lasso. We empirically show that the standard K-fold cross-validation, although very popular, is not suitable to calibrate the tuning parameter in the adaptive lasso. We propose a simple alternative cross-validation scheme, which is then shown to outperform the standard K-fold cross-validation on simulated examples.

Keywords: Causal inference, mediation analysis, analysis of high-dimensional data.

Image en couverture : Champ de coquelicots. Claude Monet.

ED 512 INFOMATHS Ecole doctorale

UNIVERSITÉ DE LYON