COLLEGE MATHS, TELECOMS
DOCTORAL INFORMATIQUE, SIGNAL
BRETAGNE SYSTEMES, ELECTRONIQUE

Université de Rennes

HABILITATION À DIRIGER DES RECHERCHES

L'UNIVERSITÉ DE RENNES

ÉCOLE DOCTORALE N° 601
Mathématiques, Télécommunications, Informatique, Signal, Systèmes, Électronique
Spécialité : *Mathématiques et leurs Interactions*

Par
Valérie Garès

Data integration, survival analysis and contributions for medical data analysis

Habilitation à diriger des recherches présentée et soutenue à l'INSA de Rennes, le 30 avril 2025
Unité de recherche : IRMAR

Rapporteur·trice·s avant soutenance :

Anne-Laure Boulesteix	Professeure (LMU, Munich, Allemagne)
Chantal Guihenneuc	Professeure (Université Paris Descartes Laboratoire MAP5, France)
Pierre Joly	Professeur (Université Victor Segalen Bordeaux 2, France)

Composition du Jury :

Président :	Jean-François Dupuy	Professeur (INSA, Rennes, France)
Examinatrice :	Anne Ruiz-Gazen	Professeure (Toulouse school of economics, France)
Examinateur :	Aurélien Latouche	Professeur (Conservatoire national des arts et métiers, Paris, France)

Remerciements.

Je souhaite tout d'abord exprimer ma reconnaissance à Anne-Laure Boulesteix, Chantal Guihenneuc et Pierre Joly qui ont accepté de rapporter ce travail et les remercier pour leur disponibilité. Je remercie également Anne-Ruiz Gazen et Aurélien Latouche qui ont accepté de faire partie du jury. Je remercie Jean-François Dupuy pour ses conseils dans l'organisation de cette soutenance.

Je tiens à remercier chaleureusement tous mes collègues de l'INSA pour leur soutien et leur collaboration. Je suis particulièrement reconnaissante envers Jean-François, Mounir, Olivier et James, qui ont veillé à ce que je ne prenne pas trop de responsabilités administratives (même si j'en ai pris de mon plein gré), et Patricia, Martine et Daviane, pour leur aide précieuse dans les tâches administratives. Mes remerciements vont également à Pierrette pour ses conseils pédagogiques avisés, à Jérémy pour notre collaboration enrichissante, et à Laurent pour notre collaboration efficace dans le cadre du séminaire entreprise. Merci à Léo (et à nouveau JaFar) pour les séances de footing partagées, ainsi qu'à Camar, Jean-Louis, Erwan, Marc, Aziz, Loïc, Rozenn, Dominique et tous les autres collègues pour les discussions conviviales à la salle café.

Je tiens à exprimer toute ma gratitude envers mes collaboratrices : Valérie, Magalie et Audrey pour son important soutien, ainsi qu'à mes collègues de Rennes 2 : Nicolas K. et Laurent pour leur bonne humeur permanente, et ceux de l'ENSAI, Guillaume et Brigitte. Je remercie également mes collègues vannetais, Nicolas C. et Chloé, ainsi que mes collègues toulousains, Nicolas S. et Philippe. Un merci particulier à mon collègue toulousain Gregory Guernec, sans qui le package `OTrecod` n'aurait pas vu le jour. Je remercie également mes collègues de l'IRSET et l'EHESP en particulier Nathalie et Nolwenn. Je tiens à réserver une place particulière à Madison et Ioana pour leur soutien et leurs précieux conseils toujours pertinents. Je souhaite remercier à nouveau Anne, qui a été mon guide depuis le début et m'a aidée à trouver ma place dans le milieu de la recherche. Les apéros de mathématiciennes ont été un précieux soutien dès mon arrivée, et pour cela, je remercie Barbara. Mes remerciements vont également aux membres du comité parité.

Merci à mes amis Toulousains, toujours fidèles : Mélanie, Janie, Lucile, Pierre, mon "ex" belle soeur Claire, les rennais : les mamans de l'école Jules Isaac (les papas aussi) pour

nos sorties sans enfant et bien-sûr Mariline, Benoît, Jérémy A., Guillaume, Arnaud, Emilie et de nouveau Nicolas K. pour leur précieux soutien et nos week-end camping, Marine, pour le soutien mutuel que nous nous sommes apporté lors de la reprise du footing après nos grossesses ! Mes dernières pensées vont à ma famille, mes parents, ma mamie, mon parrain, mon grand cousin Christophe, mon cher frère Laurent, Lucile, ma petite nièce Louise, Bernard, bien sûr Rémi et surtout, je laisse le meilleur pour la fin, mes adorables et sages enfants Eloïse et Lucas qui m'apportent un bel équilibre.

A mes enfants
Eloïse et Lucas

Table of Contents

Chapitre 1

Introduction

La pluridisciplinarité constitue un élément moteur de mes recherches. Depuis le début de ma thèse, j'interagis avec des spécialistes issus de divers domaines, notamment la santé publique (épidémiologistes, cliniciens, biostatisticiens) ainsi que des champs théoriques et/ou appliqués des statistiques. Ma démarche de recherche repose d'abord sur l'analyse d'un problème appliqué, souvent lié à une question de santé spécifique, afin de développer des méthodologies nouvelles permettant d'y répondre de manière précise. Cette approche explique la diversité de mes activités de recherche, marquées par des conversions thématiques régulières, tant sur le plan méthodologique qu'applicatif. Cela se reflète également dans la variété des concepts abordés dans les différents chapitres, ainsi que dans le nombre de collaborateurs issus de champs thématiques variés. Cette manière de pratiquer la recherche définit non seulement mon activité passée, mais oriente aussi mon approche future, car c'est la voie que je souhaite continuer à suivre. Je tiens ici à remercier profondément tous mes collaborateurs et à rappeler que ce manuscrit n'existerait pas sans eux.

Trois thématiques principales de mes travaux de recherche peuvent être distinguées. La première relève de l'intégration de données : l'appariement statistique et le chaînage de données. La deuxième concerne l'analyse de données de survie dont une partie sur l'analyse de données chaînées. La troisième partie regroupe différentes contributions en modélisation pour l'analyse de données médicales telles que l'analyse causale, longitudinale ou fonctionnelle. Mes différents travaux abordent des thématiques statistiques différentes et n'ont pas de lien spécifique bien qu'on retrouve des caractéristiques communes telles que la présence de variables latentes avec l'utilisation de l'algorithme EM pour l'estimation des paramètres du maximum de la vraisemblance. Un titre plus général pourrait être "l'augmentation ou la complétion de données" car il englobe la majorité de mes travaux, bien qu'il en laisse de côté une partie non négligeable. Une description des thèmes qui seront abordés est détaillée dans la suite de cette introduction.

1.1 Intégration de données

Mon sujet principal de recherche porte sur le développement de méthodes dédiées à l'intégration de données multi-sources utilisant la théorie du transport optimal qui se distingue en trois parties : l'appariement statistique en particulier le recodage des variables (section 1.1.1), l'adaptation de domaine (section 1.1.2) et le chaînage de données (section 1.1.3).

1.1.1 Recodage des variables

J'ai débuté ce travail sur l'appariement statistique et en particulier le recodage des variables lors de mon deuxième post-doctorat, effectué à l'INSERM [1] de Toulouse UMR 1295 CERPOP dans l'équipe EQUITY [2]. Le problème de recodage de variables peut se produire lorsqu'une variable catégorielle n'est pas codée selon la même échelle dans les deux sources de données, c'est à dire quand elle a des modalités de terminologies différentes et de nombres différents. Dans la cohorte ELFE [3], le codage correspondant à l'état de santé de la mère a été modifié entre deux vagues de recrutement. L'originalité des méthodes proposées réside dans l'utilisation de la théorie du transport optimal. Ces travaux ont été effectués en collaboration avec *Nicolas Savy* (IMT [4], Université Paul Sabatier, Toulouse), *Chloé Dimeglio* et *Gregory Guernec* (INSERM UMR 1295 CERPOP, Toulouse) [VG7].

J'ai poursuivi ce travail dès mon arrivée à l'INSA en tant que maîtresse de conférences avec la collaboration de *Jérémy Omer* (IRMAR [5], INSA [6], Rennes), chercheur en optimisation [VG2] (avec une application sur les données du NCDS [7]). Je poursuis actuellement ces travaux avec *Ioana Gavra* (IRMAR, Université de Rennes 2), *Nicolas Courty*, *Chloé Friguet* (Université de Bretagne Sud, Vannes) et *Pierre Navaro* (CNRS, IRMAR, Rennes). La problématique de recodage des variables est également présente dans les bases de données

1. Institut national de la santé et de la recherche médicale
2. Incorporation biologique, inégalités sociales, épidémiologie du cours de la vie, cancers et maladies chroniques, interventions, méthodologie
3. Étude Longitudinale Française depuis l'Enfance
4. Institut de Mathématiques de Toulouse
5. Institut de recherche mathématique de Rennes
6. Institut National des sciences appliquées
7. National Child Development Study

dont dispose l'IRSET [8], institut de santé à Rennes : les cohortes EDEN [9] et PELAGIE [10] sont des bases de données d'application pour ces travaux de recherche actuels sur ce sujet effectués en collaboration avec *Nathalie Costet* (IRSET, Rennes).

Nous avons également développé un package R (`OTrecod`) déposé sur le CRAN [11] [VG9] permettant l'accessibilité de ces travaux.

1.1.2 Adaptation de domaine hétérogène

En parallèle, depuis 2020, j'ai développé une collaboration avec l'équipe Obélix (IRISA [12]) à Vannes, *Nicolas Courty* et *Chloé Friguet*, sur une problématique proche : l'adaptation de domaine hétérogène. Dans ce cadre, nous avons encadré le stage de *Marion Jeammart*. L'objectif de ce travail est de transférer des connaissances d'un domaine source vers un domaine cible, où les deux domaines ont des distributions de données différentes afin de prédire/expliquer la même variable d'intérêt. De plus, les domaines source et cible n'ont pas seulement des distributions de données différentes, mais impliquent également des espaces des variables explicatives distincts. Ces travaux ont été présentés dans différentes conférences ($53^{ème}$ journées de la Société Française de Statistique (2022) et Journées MAS (2022)) et une publication est en cours de rédaction.

1.1.3 Chaînage de données

Cette problématique connexe a émergé à partir de discussions menées au sein d'un groupe de travail que nous avions mis en place en 2018, réunissant des chercheurs de l'EHESP [13], d'Agrocampus et de l'ENSAI [14]. Ces discussions se sont concrétisées par la mise en place de la thèse de *Huan Tanh Vo* entre différents instituts : l'EHESP, l'IRMAR et l'IRT [15] b-com. Elle a été co-encadrée avec *Guillaume Chauvet* (CREST [16], ENSAI,

8. Institut de recherche en santé, environnement et travail
9. première Étude de cohorte généraliste menée en France sur les Déterminants pré- et postnatals précoces du développement psychomoteur et de la santé de l'ENfant
10. Perturbateurs Endocriniens : Étude Longitudinale sur les Anomalies de la Grossesse, l'Infertilité et l'Enfance
11. Comprehensive R Archive Network
12. Institut de Recherche en Informatique et Systèmes Aléatoires
13. Ecole des hautes études en santé publique
14. École Nationale de la Statistique et de l'Analyse de l'Information
15. Institut de recherche technologique
16. Center for Research in Economics and Statistics

Rennes), *Andre Happe* (Université de Rennes 1) et *Stephane Paquelet* (IRT bcom). Ces travaux englobent deux problématiques :

1. le chaînage du registre GETBO [17] et des données du Système National des Données de Santé (SNDS) afin d'enrichir l'information du registre [VG14],
2. la prise en compte de l'erreur due au chaînage dans les analyses de données de survie (qui sera présentée dans la section 1.2.3 en "Analyse de données de survie") [VG13].

1.2 Analyse de données de survie

Mes travaux de thèse et de post-doctorat ont été motivés par une problématique statistique soulevée lors de l'analyse d'un essai clinique et ont principalement porté sur l'analyse de durée de vie (sections 1.2.1 et 1.2.2). La dernière partie concerne l'analyse de données de survie de données chaînées (introduite section 1.1.3).

1.2.1 Analyses de survie des essais de prévention

Au cours de ma thèse effectuée à l'INSERM UMR 1295 CERPOP à Toulouse dans l'équipe "Vieillissement et maladie d'Alzheimer" sous la supervision de *Nicolas Savy* et *Sandrine Andrieu*, j'ai travaillé sur les tests statistiques dédiés à la détection des effets tardifs d'un traitement dans le cadre d'essais cliniques de prévention de la maladie d'Alzheimer [VG6, VG3, VG4].

1.2.2 Risques compétitifs en analyse de survie

A l'issue de mon travail de thèse, j'ai effectué un post-doctorat au Centre des essais cliniques NHMRC [18] à l'université de Sydney en Australie où mes travaux se sont principalement concentrés sur l'analyse des risques compétitifs appliquée à la survie de cancers. Ces travaux ont été réalisés en collaboration avec *Malcolm Hudson* (Macquarie University, Sydney), *Val Gebski* (NHMRC, University of Sydney) et *Maurizio Manuguerra* (Macquarie University, Sydney) [VG5].

17. Groupe d'étude de la Thrombose de Bretagne Occidentale
18. National Health and Medical Research Council

1.2.3 Analyses de données de survie de données appariées

Comme évoquée dans la section 1.1.3, la seconde partie de la thèse de *Huan Vo Tanh* concernait l'analyse statistique des données appariées quand la variable d'intérêt était une durée de vie. Avec *Jean-François Dupuy* (IRMAR, INSA, Rennes) et *Samuel Bowong* (Université de Douala, Cameroun), nous poursuivons ces travaux dans le cadre de la thèse de *Vanessa Chezeu* que nous avons rencontrée lors d'une école CIMPA [19] au Sénégal. Cette thèse est en co-tutelle avec le Cameroun.

1.3 Diverses contributions pour l'analyse de données médicales

En parallèle de ces deux principaux sujets, j'ai travaillé sur différents thèmes, en particulier l'analyse causale (section 1.3.1), l'analyse de données longitudinales (section 1.3.2) et fonctionnelles (section 1.3.3).

1.3.1 Analyse causale : estimateur de variance pour le score de propension généralisé

J'ai développé une collaboration avec *David Hajage* (APHP [20], Paris) et *Guillaume Chauvet* portant sur l'estimation de variance pour le score de propension généralisé [VG1].

1.3.2 Analyse de données longitudinales avec valeurs atypiques : estimation robuste des modèles mixtes

J'ai également collaboré avec *Anne Ruiz-Gazen* (Toulouse School of economics) et *Rik Lopuhaä* (Delft University of Technology) sur l'estimation robuste pour les modèles mixtes utilisés dans le cadre de l'analyse de données longitudinales avec des valeurs atypiques [VG12].

19. Centre International de Mathématiques Pures et Appliquées
20. Assistance Publique – Hôpitaux de Paris

1.3.3 Analyse de données fonctionnelles

Je collabore actuellement avec *Valérie Monbet* (IRMAR, Université de Rennes 1) et *Madison Giacofci* (IRMAR, Université de Rennes 2) sur une nouvelle thématique : la modélisation des données fonctionnelles. En particulier, nous avons développé un modèle de mélange pour l'analyse canonique des corrélations sur données fonctionnelles.

1.4 Conseils méthodologiques pour l'épidémiologie

Durant la thèse et les deux contrats post-doctoraux, j'ai partagé ma recherche entre la statistique et l'épidémiologie. En épidémiologie, j'ai notamment participé à l'écriture de plusieurs articles scientifiques en y apportant mes compétences sur la méthodologie [VGa3, VGa1, VGa5]. J'ai également réalisé les analyses statistiques de différents articles [VGa8, VGa2]. Dans le domaine de l'**épidémiologie sociale**, en collaboration avec *Cyrille Delpierre* et *Michelle Kelly-Irving* (et coauteurs, INSERM UMR 1295 CERPOP équipe EQUITY), j'ai publié un article en premier auteur mesurant l'association entre l'environnement social précoce et l'infection par le virus Epstein Barr dans la cohorte "Millennium Cohort Study" [VGa4]. Un autre travail en collaboration avec *Emma Gibbs*, *Val Gebski* et *Karen Byth* (NHMRC, Université de Sydney), propose une méthodologie sur l'estimation d'un nombre acceptable de sujets à risques pour garantir une interprétation fiable d'une courbe de Kaplan-Meier. En effet, cette question s'est posée à l'issue de mon travail de thèse, en particulier dans le cadre de l'étude Guidage sur laquelle nous avons travaillé. En fin d'étude, des effets tardifs du traitement étaient visibles sur le graphique comparant les courbes de Kaplan-Meier des groupes traitement et placebo, alors que peu de sujets à risque restaient dans l'analyse. Ce travail a été publié dans une revue plus appliquée de santé publique afin de fournir un outil facilement exploitable par les épidémiologistes du domaine [VG8].

Avec *Jean-François Dupuy*, nous avons apporté nos compétences sur la méthodologie d'une étude portant sur l'effet de la pollution de l'air sur le risque de cancer [VGa6, VGa7] (encadrée par des chercheurs de l'IRSET, *Emeline Lequy* et *Benedicte Jacquemin*). J'ai aussi encadré plusieurs stages de recherche avec *Nathalie Costet* (IRSET) sur l'effet de la pollution de l'air sur le comportement de l'enfant, *Patrick Pladys* (CIC[21], CHU de Rennes) sur

21. Centre d'investigation clinique de Rennes

la modélisation prédictive du devenir d'enfants prématurés et *Nolwenn Le Meur* (EHESP) sur la modélisation des parcours de soins et évolutions du système de santé.

Enfin, dans le cadre de la thèse de *Abdoulaye Koroko*, je me suis intéressée aux méthodes d'optimisation basées sur le gradient naturel pour les réseaux de neurones profonds [VG11, VG10].

Dans un souci de concision, je présenterai uniquement les travaux méthodologiques en analyse de données de survie dans le chapitre dédié : les travaux sur le nombre acceptable de sujets à risques pour garantir une interprétation fiable d'une courbe de Kaplan-Meier et la méthodologie utilisée pour l'étude portant sur l'effet de la pollution de l'air. Ces deux travaux me semblent illustrer mes apports méthodologiques ainsi que les domaines d'application possibles.

1.5 Direction de mes recherches futures

Je viens d'obtenir un poste de CPJ à l'INRIA. Mon objectif pour les cinq prochaines années est de créer une équipe en science des données appliquée à la santé publique, en collaborant avec des chercheurs de l'IRSET et de l'EHESP en santé publique, ainsi qu'avec ceux de l'IRMAR et de l'IRISA. Je prévois de développer mon projet de recherche en collaboration avec mes différents partenaires autour de deux axes principaux : l'intégration de données multi-sources et l'analyse des données de spectrométrie, avec pour but d'identifier l'impact des stress environnementaux durant l'enfance sur les événements de santé. Mes différentes perspectives se focalisent autour de ces thématiques.

Pour des raisons de concision, toutes les démonstrations, la plupart des simulations et certains détails mathématiques sont omis. Ils peuvent être trouvés dans les références correspondantes.

Chapitre 2

Introduction (english)

Multidisciplinarity is a driving force in my research. Since the beginning of my PhD, I have interacted with specialists from various fields, notably public health (epidemiologists, clinicians, biostatisticians), as well as from theoretical and/or applied statistical domains. My research approach is primarily based on analyzing an applied problem, often related to a specific health question, in order to develop new methodologies that provide precise answers. This approach explains the diversity of my research activities, marked by regular thematic changes, both in methodology and application. This is also reflected in the various concepts introduced across different chapters and the number of collaborators from diverse thematic fields. This way of conducting research not only defines my past activity but also guides my future approach, as it is the path I intend to continue following. Here I want to deeply thank all my co-workers and to recall that this manuscript would not exist without them.

Three main themes can be distinguished in my research work. The first relates to data integration : statistical matching and record linkage. The second theme concerns survival data analysis, including the analysis of matched data. The third part concerns various contributions to modeling for the analysis of medical data such as causal, longitudinal, or functional analysis. My works address different statistical themes and do not have a specific link, although they share common characteristics such as the presence of latent variables with the use of the EM algorithm for maximum likelihood parameter estimation. A general title could be "data augmentation or completion" because it allows to group a majority of my work but it leaves out a significant part of it. A detailed description of the different topics is provided in this introduction.

2.1 Data integration

My main research topic focuses on the development of methods for the multi-source data integration using optimal transport theory, which can be divided into three parts : statistical matching in particular the data recoding (section 2.1.1), heterogeneous domain adaptation (section 2.1.2), and record linkage (section 2.1.3).

2.1.1 Data recoding

I initiated this work on statistical matching and variable recoding, during my second post-doctorate at INSERM UMR CERPOP within the "Biological Incorporation, Social Inequalities, Life Course Epidemiology, Cancer and Chronic Diseases, Interventions, Methodology" (EQUITY) team. The problem of variable recoding can occur when a categorical variable is not coded on the same scale in two data sources, meaning that it has different terminology and number of modalities. In the ELFE cohort, the coding that corresponds to the mother's health status was modified between two recruitment waves. The uniqueness of the proposed methods to address this issue lies in using optimal transport. These works were conducted in collaboration with *Nicolas Savy* (Paul Sabatier University, Toulouse), *Chloé Dimeglio*, and *Gregory Guernec* (INSERM U1295) [VG7].

This research continued after I joined INSA as an assistant professor, in collaboration with *Jérémy Omer* (INSA, Rennes), a researcher in optimization [VG2] (with an application to NCDS data). I continue now this research with *Ioana Gavra*, *Nicolas Courty*, *Chloé Friguet* (University of Bretagne Sud, Vannes) and *Pierre Navaro* (CNRS, IRMAR, Rennes). The variable recoding issue is also present in the databases held by IRSET, health institute at Rennes : the EDEN and PELAGIE cohorts serve as application datasets for my current research, in collaboration with *Nathalie Costet* (IRSET).

We have also developed an R package (`OTrecod`) that has been submitted to the CRAN [VG9] facilitating the accessibility of these works.

2.1.2 Heterogeneous domain adaptation

In parallel, since 2020, I have also established a collaboration with the Obélix team (IRISA, Vannes) with *Nicolas Courty* and *Chloé Friguet* on a related issue : heterogeneous

domain adaptation. In this context, we supervised *Marion Jeammart*'s internship. The objective of this work is to transfer knowledge from a source domain to a target domain, where the two domains have different data distributions in order to predict/explain the same outcome. Moreover, the source and target domains not only have different data distributions but also involve distinct explanatory variable spaces. This work has been presented at various conferences (53rd Journées de la Société Française de Statistique (2022), Journées MAS (2022)) and a publication is being written.

2.1.3 Record linkage

This related issue emerged from discussions within a working group we established in 2018, involving EHESP, Agrocampus, and ENSAI. These discussions led to *Huan Tanh Vo*'s PhD across different institutes : EHESP, IRMAR and IRT b-com. It was co-supervised by *Guillaume Chauvet* (ENSAI, Rennes), *Andre Happe* (University of Rennes 1) and *Stephane Paquelet*. This theme encompasses two issues :

1. the record linkage of clinical database GETBO with data from the French national health data system (SNDS) in order to enrich the information in the clinical database [VG14]
2. Incorporating matching error in survival data analysis (which I will present in the section 2.1.3 in "Survival Data Analysis") [VG13].

2.2 Survival analysis

My doctoral and postdoctoral works were motivated by an issue encountered in a clinical trial and focused on survival analysis (sections 2.2.1 et 2.2.3). The last part concerns the survival data analysis of matched data (introduced in section 2.1.3).

2.2.1 Survival analysis of preventive clinical trials

During my PhD conducted at INSERM UMR CERPOP in Toulouse within the "Aging and Alzheimer's Disease" team, under the supervision of *Nicolas Savy* et *Sandrine Andrieu*, I worked on statistical tests to detect late treatment effects in the context of clinical trials for Alzheimer's disease prevention [VG6, VG3, VG4].

2.2.2 Competitive risks in survival analysis

Then, I did a post-doctorate at the NHMRC Clinical Trials Centre at the university of Sydney in Australia, where my work mainly concerned competitive risk analysis applied to cancer survival. These works were carried out in collaboration with *Malcolm Hudson* (Macquarie University, Sydney), *Val Gebski* (NHMRC, University of Sydney) and Maurizio Manuguerra (Macquarie University, Sydney) [VG5].

2.2.3 Survival analysis of matched data

As mentioned in section 2.1.3, the second part of *Huan Vo Tanh*'s PhD work concerns the statistical analysis of matched data when the variable of interest is a lifetime. With *Jean-François Dupuy* (INSA, Rennes) and *Samuel Bowong* (University of Douala, Cameroun), we are continuing this work with the supervision of *Vanessa Chezeu*'s PhD, in cotutelle with Cameroon, whom we met during a CIMPA school in Senegal.

2.3 Some contributions to the analysis of medical data

In parallel with these two main topics, I could work on different subjects in particular causal analysis (section 2.3.1), longitudinal data analysis (section 2.3.2), and functional data analysis (section 2.3.3).

2.3.1 Causal analysis : variance estimator for the generalized propensity score

I worked with *David Hajage* (APHP, Paris) and *Guillaume Chauvet* on variance estimation for generalized propensity score [VG1].

2.3.2 Longitudinal data analysis with outliers : robust estimation of mixed models

I also developed a collaboration with *Anne Ruiz-Gazen* (Toulouse School of Economics) and *Rik Lopuhaä* (Delft University of Technology) on robust estimation for mixed models used in longitudinal data analysis with outliers [VG12].

2.3.3 Functional data analysis

I am currently collaborating with *Valérie Monbet* (IRMAR, University of Rennes 1) and *Madison Giacofci* (IRMAR, University of Rennes 2) on a new theme : the modeling of functional data. In particular, we have developed a mixture model for canonical correlation analysis for functional data.

2.4 Some methodological guidelines for epidemiology

Throughout the PhD and the two post-doctoral positions, my research has been divided between two frameworks : statistics and epidemiology. In epidemiology, I collaborated on certain articles providing my expertise in methodology [VGa3, VGa1, VGa5]. I realized statistical analysis of different articles [VGa8, VGa2]. In the field of social epidemiology, with *Cyrille Delpierre* and *Michelle Kelly-Irving* (and co-authors from INSERM UMR CERPOP EQUITY team), I published a first-author article on the association between early social environment and Epstein-Barr virus infection in the Millennium Cohort Study cohort [VGa4]. Another work with *Emma Gibbs*, *Val Gebski*, and *Karen Byth* (NHMRC, University of Sydney) proposed a methodology on the estimation of the acceptable number at-risk subjects to interpret a Kaplan-Meier curve. This question arose from my thesis work, specifically in the context of the Guidage study on which we worked. At the end of the study, late treatment effects were visible on the graph comparing the Kaplan-Meier curves of the treatment and placebo groups, with few subjects at risk remaining in the analysis. This article was chosen to be published in an epidemiology journal to provide a readily usable tool for epidemiologists [VG8].

With *Jean-François Dupuy*, we have also provided methodological advice for a study on the effect of air pollution on cancer risk [VGa6, VGa7] (with researchers from IRSET, *Emeline Lequy*, and *Benedicte Jacquemin*). I have also supervised several research internships with *Nathalie Costet* (IRSET) on the effect of air pollution on children's behavior, *Patrick Pladys* (CIC, CHU of Rennes) on predictive modeling of premature infant future, and *Nolwenn Le Meur* (EHESP) on the modeling of care pathways and developments in the healthcare system.

With the supervision of *Abdoulaye Koroko*'s PhD, I have learnt on optimization methods

based on natural gradient for deep neural networks [VG11, VG10].

In the sake of conciseness, I will present only the works on survival data analysis in the dedicated chapter : the research on the acceptable number at-risk subjects to interpret a Kaplan-Meier curve, and the methodology used for the study on the effect of air pollution. Indeed, these two works seem to illustrate my methodological contributions as well as the possible areas of application.

2.5 Direction of my future research

I have just obtained a CPJ position at INRIA. My objective for the next five years is to create a team in data science applied to public health, collaborating with researchers from IRSET and EHESP in public health, as well as from IRMAR and IRISA in data sciences. I plan to develop my research project in collaboration with my various partners along two main axes : multi-source data integration and the analysis of spectrometry data, to measure the impact of environmental stressors during childhood on health outcomes. My various perspectives are focused on these themes.

In a sake of compactness all the different proofs, most of the simulations and some mathematical details are omitted. They could be found in the corresponding references.

ORDERED LIST OF MY PUBLICATIONS IN STATISTICS

[VG1] V. GARÈS, G. CHAUVET et D. HAJAGE, « Variance estimators for weighted and stratified linear dose-response function estimators using generalized propensity score », in : *Biometrical Journal* 64.*1* (2022), p. 33-56.

[VG2] V. GARÈS et J. OMER, « Regularized optimal transport of covariates and outcomes in data recoding », in : *Journal of the American Statistical Association* (2020), p. 1-14.

[VG3] V. GARÈS et al., « A comparison of the constant piecewise weighted logrank and Fleming-Harrington tests », in : *Electronic Journal of Statistics* 8.*1* (2014), p. 841 -860.

[VG4] V. GARÈS et al., « An omnibus test for several hazard alternatives in prevention randomized controlled clinical trials », in : *Statistics in medicine* 34.*4* (2015), p. 541-557.

[VG5] V. GARÈS et al., « Effect of a correlated competing risk on marginal survival estimation in an accelerated failure time model », in : *Communications in Statistics - Simulation and Computation* (2024).

[VG6] V. GARÈS et al., « On the Fleming-Harrington test for late effects in prevention randomized controlled trials », in : *Journal of Statistical Theory and Practice* 11 (2017), p. 418-435.

[VG7] V. GARÈS et al., « On the use of optimal transportation theory to recode variables and application to database merging », in : *The international journal of biostatistics* 16.*1* (2019).

[VG8] V. GEBSKI et al., « Data maturity and follow-up in time-to-event analyses », in : *International journal of epidemiology* 47.*3* (2018), p. 850-859.

[VG9] G. Guernec et al., « The R Journal : OTrecod : An R Package for Data Fusion using Optimal Transportation Theory », in : *The R Journal* 14 (4 2023), p. 195-222, issn : 2073-4859.

[VG10] A. Koroko et al., « Analysis and comparison of two-level KFAC methods for training deep neural networks », in : *Optimization Methods and Software* (2024), p. 1-33.

[VG11] A. Koroko et al., « Efficient approximations of the fisher matrix in neural networks using kronecker product singular value decomposition », in : *arXiv preprint arXiv :2201.10285* (2022).

[VG12] H. P. Lopuhaä, V. Garès et A. Ruiz-Gazen, « S-estimation in Linear Models with Structured Covariance Matrices », in : *Annals of statistics* 51.*6* (2023), p. 2415-2439.

[VG13] T. H. Vo et al., « Cox regression with linked data », in : *Statistics in Medicine* (2022).

[VG14] T. H. Vo et al., « Extending the Fellegi-Sunter record linkage model for mixed-type data with application to the French national health data system », in : *Computational Statistics & Data Analysis* 179 (2023), p. 107656.

ORDERED LIST OF MY PUBLICATIONS IN JOURNALS APPLIED TO MEDICAL

[VGa1] R. CASTAGNÉ et al., « Allostatic load and subsequent all-cause mortality : which biological markers drive the relationship ? Findings from a UK birth cohort », in : *European journal of epidemiology* 33 (2018), p. 441-458.

[VGa2] N. COLEY et al., « A longitudinal study of transitions between informal and formal care in Alzheimer disease using multistate models in the European IC-TUS cohort », in : *Journal of the American Medical Directors Association* 16.*12* (2015), 1104-e1.

[VGa3] D. W. GARSED et al., « Homologous recombination DNA repair pathway disruption and retinoblastoma protein loss are associated with exceptional survival in high-grade serous ovarian cancer », in : *Clinical Cancer Research* 24.*3* (2018), p. 569-580.

[VGa4] V. GARÈS et al., « The role of the early social environment on Epstein Barr virus infection : a prospective observational design using the Millennium Cohort Study », in : *Epidemiology & Infection* 145.*16* (2017), p. 3405-3412.

[VGa5] A. GUILLEN et al., « Suicidal ideation and traumatic exposure should not be neglected in epileptic patients : a multidimensional comparison of the psychiatric profile of patients suffering from epilepsy and patients suffering from psychogenic nonepileptic seizures », in : *Frontiers in Psychiatry* 10 (2019), p. 303.

[VGa6] E. LEQUY et al., « Contribution of long-term exposure to outdoor black carbon to the carcinogenicity of air pollution : Evidence regarding risk of cancer in the gazel cohort », in : *Environmental health perspectives* 129.*3* (2021), p. 037005.

[VGa7] E. LEQUY et al., « Influence of exposure assessment methods on associations between long-term exposure to outdoor fine particulate matter and risk of cancer in the French cohort Gazel », in : *Science of the Total Environment* 820 (2022), p. 153098.

[VGa8] M. Soto et al., « Living Alone with Alzheimer's Disease and the Risk of Adverse Outcomes : Results from the Plan de Soin et d'Aide dans la maladie d'Alzheimer Study », in : *Journal of the American Geriatrics Society* 63.*4* (2015), p. 651-658.

Chapitre 3

Data integration

3.1 General introduction on data integration

My main research topic focuses on the development of methods for the multi-source data integration, which can be divided into three parts : statistical matching in particular the data recoding and domain adaptation (section 3.2), and record linkage (section 3.3).

In these works, we focus on two objectives resulting from the integration of databases :

- The first objective is to obtain a dataset enriched in terms of individuals, thus having more realizations of the random variables. However, harmonizing the variables in both databases may be necessary. In our work on variable recoding, we address the case where a categorical variable recording the same information is coded differently in the two databases. This means the categorical variable has different terminologies and a different number of categories. In our work on heterogeneous domain adaptation, the explanatory variables may be in different spaces in the two databases. This objective is illustrated in Figure 3.1. Statistical matching or domain adaptation corresponds to the set of methods aimed at achieving this objective.

- A second objective is the record linkage which aims to enrich the information at the individual level in a database, meaning adding variables to the database that were originally recorded in another database. It is common to enrich a database by using data from the SNDS, which covers the entire French population. This objective is illustrated in Figure 3.2.

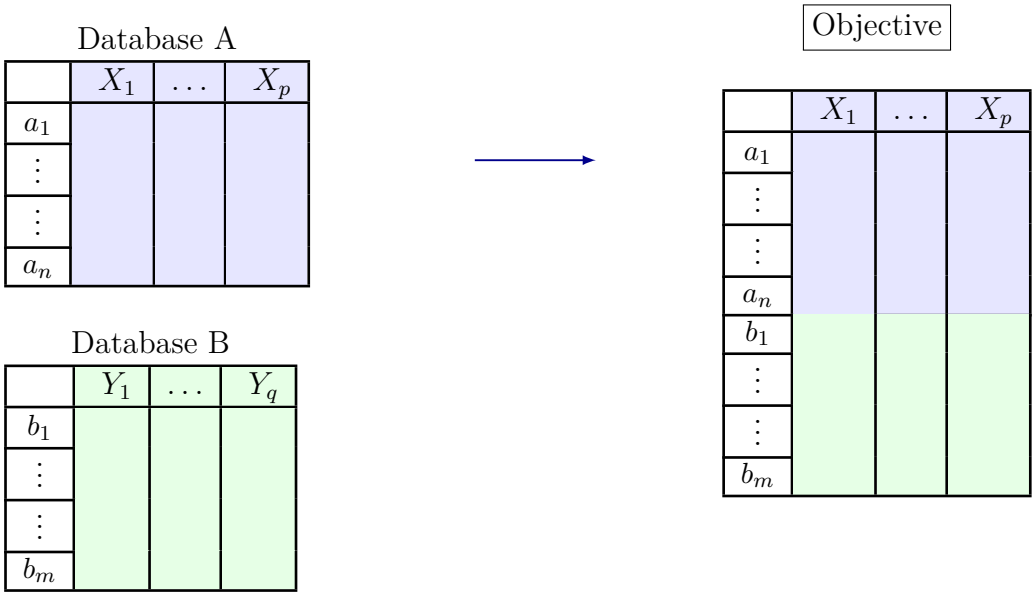

FIGURE 3.1 – Statistical matching

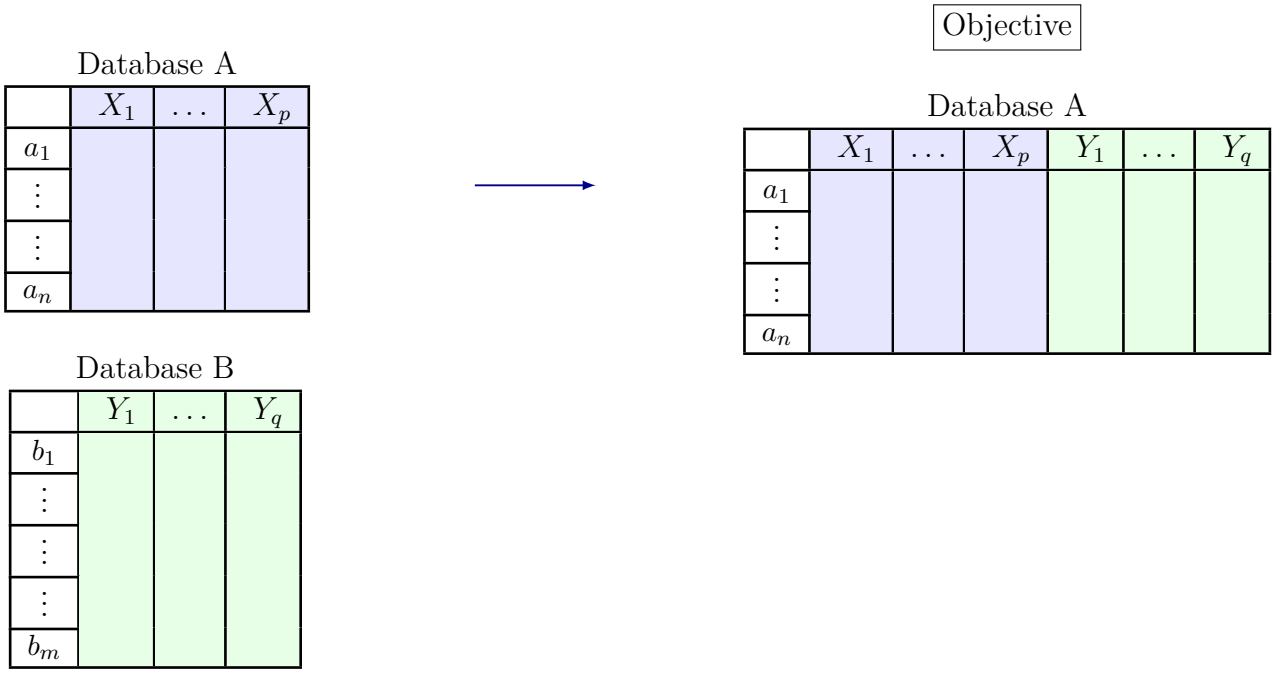

FIGURE 3.2 – Record linkage

3.2 Variable recoding problem and heterogeneous domain adaptation using optimal transport theory

Motivations :

- **Statistical matching** : set of methods aimed at integrating two (or more) data sources with different samples referring to the same population.
  Variable recoding problem : when a categorical variable is not coded on the same scale in the two data sources.
- **Domain adaptation** : set of techniques proposed to transfer knowledge from a source domain to a target domain, where the two domains have different data distributions to explain/predict the same outcome.
  Heterogeneous Domain Adaptation : particular case of domain adaptation that deals with situations where the source and target domains not only have different data distributions but also involve different spaces of explanatory variables (features).

Contributions of the Chapter :

- Use of optimal transport theory for data recoding problem and Heterogeneous Domain Adaptation.

Collaborations :

- Data recoding problem
  - ✓ *Nicolas Savy* (IMT, Paul Sabatier University, Toulouse), *Chloé Dimeglio*, and *Gregory Guernec* (INSERM UMR 1295 CERPOP, Toulouse) [VG7], section 3.2.2.
  - ✓ *Jérémy Omer* (INSA Rennes) [VG2], section 3.2.2.
  - ✓ *Nicolas Courty*, *Chloé Friguet* (IRISA, University of Bretagne Sud, Vannes), *Ioana Gavra* (IRMAR, University of Rennes 2), and *Pierre Navaro* (CNRS, IRMAR, Rennes), section 3.2.2.
- Heterogeneous domain adaptation
  - ✓ *Nicolas Courty*, *Chloé Friguet*, *Ioana Gavra*, section 3.2.3.

Package R : `OTrecod` [VG9] for data recoding

Perspectives :

- Heterogeneous Domain Adaptation with mixed type explanatory variables.
- Variable recoding with different databases containing information collected at different times for the same individuals.
- Statistical modeling on imputed data after data integration.

3.2.1 Introduction

Problems and objectives

Variable recoding.

Problem. The problem of variable recoding arises when a variable is not coded on the same scale in the two data sources, that is to say when it has different terminologies and number of modalities. As illustrated in Figure 3.3, the problem can be formalized in terms of two data sources, A containing n_A units and B containing n_B units. It is assumed that A and B contain disjoint units. A and B share a subset of variables $\boldsymbol{X}$ (referred to as explanatory variables), and at the same time, each data source A and B distinctly observe other subsets of variables : Y in A and Z in B (outcomes), which are explained by $\boldsymbol{X}$.

A	$\boldsymbol{X} \in \mathbb{R}^p$	$Y \in \mathbb{R}$	$Z \in \mathbb{R}$
1	Observed	Observed	Non observed
⋮			
⋮			
n_A			

B	$\boldsymbol{X} \in \mathbb{R}^p$	$Y \in \mathbb{R}$	$Z \in \mathbb{R}$
1	Observed	Non observed	Non observed
⋮			
⋮			
n_B			

(a) Database A and B

	$\boldsymbol{X} \in \mathbb{R}^p$	$Y \in \mathbb{R}$	$Z \in \mathbb{R}$
1	Observed	Observed	Predict
⋮			
⋮			
n_A			
$n_A + 1$		Predict	Observed
⋮			
⋮			
$n_A + n_B$			

(b) Synthetic Dataset

FIGURE 3.3 – Statistical matching between databases A and B, including common variables (explanatory variables $\boldsymbol{X}$) and distinct variables (outcomes Y and Z respectively).

An application (from the EQUITY team at INSERM UMR CERPOP) is the French cohort study ELFE. The objective of this study is to explain how various contextual factors (such as perinatal conditions and environment) impact the development of children's

health and well-being over time and into adulthood. The variable of interest is the response to the question asked to the mother : "How would you rate your general health ?". During the first wave of data collection (january to april 2011), an ordinal scale was used with five possible responses : "excellent", "very well", "well,", "fair" and "bad". In the second wave of data collection (may to december 2011), another ordinal scale with five possible responses was used : "very well," "well", "medium", bad" and "very bad".

Objective. The data recoding problem consists in the prediction of Z_i, $i = 1, \ldots, n_A$ (or the prediction of Y_j, $j = 1, \ldots, n_B$) from the observations $(\boldsymbol{x}_i, y_i)$, $i = 1, \ldots, n_A$ and $(\boldsymbol{x}_j, z_j)$, $j = 1, \ldots, n_B$.

Heterogeneous Domain Adaptation.

Problem. Transfer learning involves training a model on a dataset explaining a continuous or categorical outcome, called the source domain, to predict the unobserved outcome in a new dataset, called the target domain. However, there may exist differences in the conditional and/or marginal distributions of variables between the source and target domains, and the learned models cannot be directly applied to predict outcomes in the target domain. Domain adaptation (DA) refers to a set of techniques proposed to overcome this issue [94]. In general, most existing domain adaptation methods assume that data from the source and target domains are represented in the same feature space, with identical dimensions. However, in some applications, this is not the case, feature spaces can be heterogeneous. This is referred to as heterogeneous domain adaptation (HDA). In the context of omics data, the explanatory variables may be the pollutants measured in the child's umbilical cord in one database S, and the same pollutants recorded in the mother's blood analysis in another database T. As illustrated in Figure 3.4, the problem can be formalized in terms of two databases : the source data S and the target data T, with respectively n_S and n_T observations. $\boldsymbol{X}^S$ and $\boldsymbol{X}^T$ may have different distributions and may be in a different feature space.

Objectives. The objective is to propose prediction models adapted to HDA. Let's denote the set of the sample indices where the outcome Y is observed in the source and target samples respectively by $\mathcal{I}^S$ and $\mathcal{I}^T$. We have concentrated on three types of HDA scenarios (see table 3.1) :

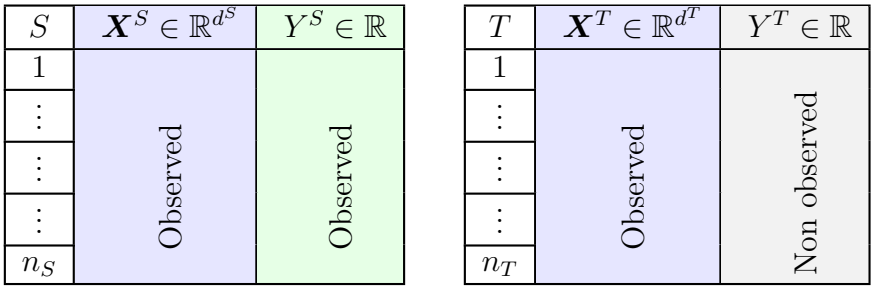

FIGURE 3.4 – Heterogeneous domain adaptation between S (source) and T (target) databases

1. *Unsupervised HDA*, where all outcomes from the source domain are observed, but none in the target domain. The objective is to train a predictive function f^T using the observations in data source S, $(\boldsymbol{x}_i^S, y_i^S)$, $i \in \mathcal{I}^S = \{1, \ldots, n_S\}$ to predict $Y^T = f^T(\boldsymbol{X}^T)$ in the target domain T.

2. *Semi-supervised HDA*, where all outcomes from the source domain and some outcomes from the target domain are observed. The objective is to train a predictive function f^T using the observations in data source S, $(\boldsymbol{x}_i^S, y_i^S)$, $i \in \mathcal{I}^S = \{1, \ldots, n_S\}$ and data target T, $(\boldsymbol{x}_j^T, y_j^T)$, $j \in \mathcal{I}^T$ to predict $Y^T = f^T(\boldsymbol{X}^T)$ in the target domain T.

3. *Partial HDA*, where outcomes from both the source and target domains are partially observed. The objective is to train two predictive functions f^T and f^S using the observations in data source S, $(\boldsymbol{x}_i^S, y_i^S)$, $i \in \mathcal{I}^S$ and data target T, $(\boldsymbol{x}_j^T, y_j^T)$, $j \in \mathcal{I}^T$ to predict $Y^T = f^T(\boldsymbol{X}^T)$ in the target domain T and to predict $Y^S = f^S(\boldsymbol{X}^S)$ in the source domain S.

	Y^S	Y^T
Unsupervised HDA	observed	non observed
Semi-supervised HDA	observed	partially observed
Partial HDA	partially observed	partially observed

TABLE 3.1 – Multiple learning cases based on the availability of outcomes Y in S and/or T databases. *Note : In the context of partial HDA, since outcomes from both domains are partially observed, the source and target domains can be interchangeable.*

Literature

Statistical matching. Classically, methods for classification learning [110] first estimate the distribution of Y given explanatory variables using the observations in dataset A $(\boldsymbol{x}_i, y_i)$, $i = 1, \ldots, n_A$ in order to predict the missing variable Y_j, $j = 1, \ldots, n_B$ in dataset B. Then, variable recoding can be seen as a missing data problem. In this context, the missing value depends only on the database to which the subject belongs and not on the variable itself. The missingness mechanism can then be considered as *missing at random*. This problem has been widely studied in the literature [78] and many existing methods for treating missing data could be used. Moreover, Y and Z refer to the same information, which can be interpreted as a latent variable. Methods of prediction of this latent class could also be applied (e.g. *class latent analysis* or *trait latent analysis* [6, 101]). The major drawback of such approaches is that the resulting methods use the information contained in each database in two independent steps that cannot capture the interrelations between them. The outcome Z, observed in B but not in A, is considered as additional information that can be incorporated into the model to improve the estimation of Y in B. Furthermore, differences between the joint distributions of $(\boldsymbol{X}, Y, Z)$ in databases A and B may exist.

Domain adaptation. Different reasons characterize the inequality between the joint distributions in source and target domains depending on the assumptions made about the conditional and marginal distributions. Under the covariate shift assumption, the differences between the database distributions are characterized by a change in the covariate distributions $\mu^{\boldsymbol{X}}$, while the conditional distributions $\mu^{Y|\boldsymbol{X}}$ remain unchanged across domains [27]. Under the target shift assumption, the differences between the databases are characterized by a change in the target distribution μ^{Y} while the conditional distribution $\mu^{\boldsymbol{X}|Y}$ is conserved across domains [94]. Optimal transport (OT) [112] has been proposed to solve DA issues [26], under target shift [94] or covariate shift assumptions [28].

Heterogeneous domain adaptation. Main methods in HDA can be classified regarding two strategies : (1) project both feature spaces into a common subspace by jointly learning the common subspace and a classifier, then iteratively align the discriminative dimensions ([120] or [16]), and (2) jointly perform implicit data reconstruction, by a source feature space transformation to align it with the target feature space, and learn a classifier [77]. See [34] for a survey of several methods of Heterogenous Transfer Learning. OT has been

also proposed to solve HDA issue. The problem can then be addressed by an algorithm, named COOT, which stands for CO-Optimal Transport [106]. The feature vectors in the source and target domains can have different dimensions, but all components of the vector must be defined in the same probability space. However, this algorithm does not allow for the prediction of a target variable.

Optimal transport background

The OT problem was originally stated by [85] and consists in finding the cheapest way to transport a pile of sand to fill a hole. Formally the problem writes as follows : consider two (Radon) spaces $\mathcal{X}$ and $\mathcal{Y}$, μ^X a probability measure on $\mathcal{X}$, and μ^Y a probability measure on $\mathcal{Y}$ and c a Borel-measurable function from $\mathcal{X} \times \mathcal{Y}$ to $[0, \infty]$. The Kantorovich's formulation of the optimal transport problem [69] consists in finding a measure $\gamma \in \Gamma(\mu^X, \mu^Y)$ that realizes the infimum :

$$\inf \left\{ \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\, \mathrm{d}\gamma(x,y) \middle| \gamma \in \Gamma(\mu^X, \mu^Y) \right\}, \tag{3.1}$$

where $\Gamma(\mu^X, \mu^Y)$ is the set of measures on $\mathcal{X} \times \mathcal{Y}$ with marginals μ^X on $\mathcal{X}$ and μ^Y on $\mathcal{Y}$.

Consider two sets of weighted samples :

$$\{(\boldsymbol{x}_i, w_i),\ i = 1 \dots n,\ \sum_{i=1}^{n} w_i = 1\} \quad \text{and} \quad \{(\boldsymbol{y}_j, v_j),\ j = 1 \dots m,\ \sum_{j=1}^{m} v_j = 1\}$$

The Kantorovich formulation given in equation (3.1) applied to these empirical measures consists in finding a coupling matrix $\boldsymbol{\pi} \in \mathbb{R}_+^{n \times m}$ that satisfies :

$$\boldsymbol{\pi}^\star = \underset{\boldsymbol{\pi} \in \Pi(\boldsymbol{w}, \boldsymbol{v})}{\arg\min} \sum_{i,j} C_{i,j} \pi_{i,j}, \tag{3.2}$$

with $\boldsymbol{C}$ a cost-matrix that defines the transport cost between two samples $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$, and $\Pi(\boldsymbol{w}, \boldsymbol{v})$ is the set of matrices $\boldsymbol{\pi} \in \mathbb{R}_+^{n \times m}$ which verifies :

$$\begin{cases} \sum_{i=1}^{n} \pi_{i,j} = v_j,\ \forall j = 1 \dots m, \\ \sum_{j=1}^{m} \pi_{i,j} = w_i,\ \forall i = 1 \dots n, \end{cases}$$

to ensure that all samples are transported. We define $\boldsymbol{\pi}^\star := \mathrm{OT}(\boldsymbol{w}, \boldsymbol{v}, \boldsymbol{C})$.

OT is an optimization problem that allows to define a distance between two probability measures, called the Wasserstein distance, which is actually used as a loss in many optimization problems (see [91] for details).

I will introduce the different algorithms that we proposed for data recoding and HDA in the next sections.

3.2.2 Variable recoding problem

Formal statement of the data recoding problem

Let A and B be two independent databases corresponding to two sets of subjects. For more concise notations, we assume without loss of generality that the two databases have equal sizes, so that they can be written as $A = \{i_1, \ldots, i_n\}$ and $B = \{j_1, \ldots, j_n\}$. Let $((\boldsymbol{X}_i, Y_i, Z_i))_{i\in A}$ and $((\boldsymbol{X}_j, Y_j, Z_j))_{j\in B}$ be two sequences of i.i.d. discrete random variables with values in $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$, where $\mathcal{X}$ is a finite subset of $\mathbb{R}^p$, and $\mathcal{Y}$ and $\mathcal{Z}$ are finite subsets of $\mathbb{R}$. Variables $(\boldsymbol{X}_i, Y_i, Z_i), i \in A$, are i.i.d copies of $(\boldsymbol{X}^A, Y^A, Z^A)$ and $(\boldsymbol{X}_j, Y_j, Z_j), j \in B$, are i.i.d copies of $(\boldsymbol{X}^B, Y^B, Z^B)$. By independence of the databases, we moreover assume that $(\boldsymbol{X}^A, Y^A, Z^A)$ are independent of $(\boldsymbol{X}^B, Y^B, Z^B)$. Every random variable is defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Finally, we assume that for all $\boldsymbol{x} \in \mathcal{X}$ the probability distributions of Y^A and Z^A given that $\boldsymbol{X}^A = \boldsymbol{x}$ are respectively equal to those of Y^B and Z^B given that $\boldsymbol{X}^B = \boldsymbol{x}$, i.e.,

Assumption 3.1.

$$\mathbb{P}(Y^A = y \mid \boldsymbol{X}^A = \boldsymbol{x}) = \mathbb{P}(Y^B = y \mid \boldsymbol{X}^B = \boldsymbol{x}),\ \forall \boldsymbol{x} \in \mathcal{X}, \forall y \in \mathcal{Y},\ \textit{and}$$
$$\mathbb{P}(Z^A = z \mid \boldsymbol{X}^A = \boldsymbol{x}) = \mathbb{P}(Z^B = z \mid \boldsymbol{X}^B = \boldsymbol{x}),\ \forall \boldsymbol{x} \in \mathcal{X}, \forall z \in Z.$$

In particular, assumption 3.1 implicitly states that $\mathbb{P}(Y^A = y \mid \boldsymbol{X}^A = \boldsymbol{x})$ is defined if and only if $\mathbb{P}(Y^B = y \mid \boldsymbol{X}^B = \boldsymbol{x})$ is defined, i.e., $\mathbb{P}(\boldsymbol{X}^A = \boldsymbol{x}) = 0$ if and only if $\mathbb{P}(\boldsymbol{X}^B = \boldsymbol{x}) = 0$. Without loss of generality, we thus restrict the domain of the explanatory variables $\mathcal{X}$ to the values, $\boldsymbol{x}$, such that $\mathbb{P}(\boldsymbol{X}^A = \boldsymbol{x}) \neq 0$ and $\mathbb{P}(\boldsymbol{X}^B = \boldsymbol{x}) \neq 0$.

The data recoding problem consists in the prediction of $(Z_i)_{i\in A}$ from independent realizations of random variables $\boldsymbol{X}$ and Y in A, $((\boldsymbol{x}_i, y_i))_{i\in A}$ and of $\boldsymbol{X}$ and Z in B,, $((\boldsymbol{x}_j, z_j))_{j\in B}$.

To be more specific, the problem is to propose an estimator of the distribution of Z^A conditional to $\boldsymbol{X}^A = \boldsymbol{x}$ and $Y^A = y$, i.e.,

$$\left\{\mathbb{P}(Z^A = z|\boldsymbol{X}^A = \boldsymbol{x}, Y^A = y), z \in \mathcal{Z}\right\}, \boldsymbol{x} \in \mathcal{X}, y \in \mathcal{Y},$$

from the observations of $\boldsymbol{X}$ and Y in A, $((\boldsymbol{x}_i, y_i))_{i\in A}$, and of $\boldsymbol{X}$ and Z in B, $((\boldsymbol{x}_j, z_j))_{j\in B}$.

Observe that since Y and Z are never jointly observed for the same subject, the assumption on conditional distributions 3.1 cannot be tested from the data.

Distributions and estimators

The discrete probability distribution of a discrete random vector V with values in $\mathcal{V}$ is given by

$$\mu^V = \sum_{v\in\mathcal{V}} \mu_v^V \delta_v,$$

where δ_v is the Dirac delta measure centered at v. If $\mathcal{V}$ is finite with cardinality $|V|$, μ^V will also refer to the vector of weights $(\mu_v^V)_{v\in\mathcal{V}}$.

Vectors $((\boldsymbol{x}_i, y_i))_{i\in A}$ and $((\boldsymbol{x}_j, z_j))_{j\in B}$ are realizations of two n samples of random variables $((\boldsymbol{X}_i, Y_i))_{i\in A}$ and $((\boldsymbol{X}_j, Y_j))_{j\in B}$ with unknown joint distribution $\mu^{(\boldsymbol{X}^A, Y^A)}$ and $\mu^{(\boldsymbol{X}^B, Z^B)}$. As a consequence, we will consider their unbiased empirical estimators given by

$$\widehat{\mu}_{n,\boldsymbol{x},y}^{(\boldsymbol{X}^A,Y^A)} = \frac{1}{n}\sum_{i\in A} \mathbb{1}_{\{\boldsymbol{X}_i=\boldsymbol{x}, Y_i=y\}}, \ \forall \boldsymbol{x} \in \mathcal{X}, y \in \mathcal{Y}, \tag{3.3}$$

$$\widehat{\mu}_{n,\boldsymbol{x},z}^{(\boldsymbol{X}^B,Z^B)} = \frac{1}{n}\sum_{j\in B} \mathbb{1}_{\{\boldsymbol{X}_j=\boldsymbol{x}, Z_j=z\}}, \ \forall \boldsymbol{x} \in \mathcal{X}, z \in \mathcal{Z}. \tag{3.4}$$

The strong law of large numbers applied to the sequences of i.i.d. random variables $((\boldsymbol{X}_i, Y_i))_{i\in A}$ and $((\boldsymbol{X}_j, Z_j))_{j\in B}$ directly yields that

$$\widehat{\mu}_{n,\boldsymbol{x},y}^{(\boldsymbol{X}^A,Y^A)} \xrightarrow[n\to+\infty]{a.s.} \mu_{\boldsymbol{x},y}^{(\boldsymbol{X}^A,Y^A)}, \forall \boldsymbol{x} \in \mathcal{X}, \forall y \in \mathcal{Y}$$
$$\widehat{\mu}_{n,\boldsymbol{x},z}^{(\boldsymbol{X}^B,Z^B)} \xrightarrow[n\to+\infty]{a.s.} \mu_{\boldsymbol{x},z}^{(\boldsymbol{X}^B,Z^B)}, \forall \boldsymbol{x} \in \mathcal{X}, \forall z \in \mathcal{Z}.$$

In what follows, we will also need an estimator of $\mu^{(\boldsymbol{X}^A,Z^A)}$. Let $\boldsymbol{x} \in \mathcal{X}$ ($\mathbb{P}(\boldsymbol{X}^B = \boldsymbol{x}) \neq 0$)

and $z \in \mathcal{Z}$, then

$$\begin{aligned}\mathbb{P}(\boldsymbol{X}^A = \boldsymbol{x}, Z^A = z) &= \mathbb{P}(Z^A = z \mid \boldsymbol{X}^A = \boldsymbol{x})\mathbb{P}(\boldsymbol{X}^A = \boldsymbol{x}) \\ &= \mathbb{P}(Z^B = z \mid \boldsymbol{X}^B = \boldsymbol{x})\mathbb{P}(\boldsymbol{X}^A = \boldsymbol{x}) \\ &= \mathbb{P}(\boldsymbol{X}^B = \boldsymbol{x}, Z^B = z)\frac{\mathbb{P}(\boldsymbol{X}^A = \boldsymbol{x})}{\mathbb{P}(\boldsymbol{X}^B = \boldsymbol{x})},\end{aligned}$$

where the second equality is a direct application of the assumption (3.1). Denoting as $\widehat{\mu}_n^{\boldsymbol{X}^A}$ and $\widehat{\mu}_n^{\boldsymbol{X}^B}$ the unbiased empirical estimators of $\mu^{\boldsymbol{X}^A}$ and $\mu^{\boldsymbol{X}^B}$, we then consider the estimator of $\mu^{(\boldsymbol{X}^A, Z^A)}$ given by, $\forall \boldsymbol{x} \in \mathcal{X}$, $\forall z \in Z$,

$$\widehat{\mu}_{n,x,z}^{(\boldsymbol{X}^A, Z^A)} = \begin{cases} \frac{\widehat{\mu}_{n,\boldsymbol{x},z}^{(\boldsymbol{X}^B, Z^B)} \widehat{\mu}_{n,\boldsymbol{x}}^{\boldsymbol{X}^A}}{\widehat{\mu}_{n,\boldsymbol{x}}^{\boldsymbol{X}^B}}, & \text{if } \widehat{\mu}_{n,\boldsymbol{x}}^{\boldsymbol{X}^B} \neq 0, \\ 0, & \text{if } \widehat{\mu}_{n,\boldsymbol{x}}^{\boldsymbol{X}^B} = 0. \end{cases} \tag{3.5}$$

The almost sure convergence of all the estimators in the expression of $\widehat{\mu}_{n,\boldsymbol{x},z}^{(\boldsymbol{X}^A, Z^A)}$ yields

$$\widehat{\mu}_{n,\boldsymbol{x},z}^{(\boldsymbol{X}^A, Z^A)} \xrightarrow[n \to +\infty]{a.s.} \mu_{\boldsymbol{x},z}^{(\boldsymbol{X}^A, Z^A)}, \forall \boldsymbol{x} \in \mathcal{X}, \forall z \in \mathcal{Z}.$$

Remark 3.1 (Estimators for comparable populations)**.** *Observe that if we add the assumption that $\boldsymbol{X}^A$ and $\boldsymbol{X}^B$ are copies of the same random variable $\boldsymbol{X}$, we immediately see that in such case, we can use the unbiased estimator $\widehat{\mu}_n^{(\boldsymbol{X}^A, Z^A)} = \widehat{\mu}_n^{(\boldsymbol{X}^B, Z^B)}$.*

Optimal transport of outcomes within data sources

Model. The first OT approach for data recoding is described in [VG7]. This version makes the additional assumption that :

Assumption 3.2. *Y^A and Z^A respectively follow the same distribution as Y^B and Z^B.*

In this setting, we aim at solving the OT problem (3.1) that pushes μ^{Y^A} forward to μ^{Z^A}. As a consequence, variable γ of (3.1) is a discrete measure with marginals μ^{Y^A} and μ^{Z^A}, represented by a $|\mathcal{Y}| \times |\mathcal{Z}|$ matrix. The cost, denoted as c is a $|\mathcal{Y}| \times |\mathcal{Z}|$ matrix, $(c_{y,z})_{y \in \mathcal{Y}, z \in \mathcal{Z}}$. To be specific, the goal is the identification of

$$\boldsymbol{\gamma}^\star \in \operatorname{argmin}_{\gamma \in \mathbb{R}_+^{|\mathcal{Y}| \times |\mathcal{Z}|}} \left\{ \langle \boldsymbol{\gamma} | \boldsymbol{c} \rangle : \boldsymbol{\gamma} \mathbf{1}_{|\mathcal{Z}|} = \mu^{Y^A}, \boldsymbol{\gamma}^T \mathbf{1}_{|\mathcal{Y}|} = \mu^{Z^A} \right\}, \tag{3.6}$$

where $\langle \cdot | \cdot \rangle$ is the dot product, $\mathbf{1}$ is a vector of ones with appropriate dimension and M^T is

the transpose of matrix M. The cost function $c_{y,z}$ measures the average distance between the explanatory variables of subjects of A satisfying $Y = y$ and subjects of B satisfying $Z = z$, that is

$$c_{y,z} = \mathbb{E}\left[d(\boldsymbol{X}^A, \boldsymbol{X}^B) \mid Y^A = y, Z^B = z\right], \tag{3.7}$$

where d can be any distance function defined on $\mathcal{X} \times \mathcal{X}$. This choice allows for a clear connection between the structures observed in databases A and B.

Remark 3.2. *Here a Hamming distance from the associated complete disjunctive tables is used but other distances are adapted to ordinal categorical variables : Spearman, Chebyshev, Kendall, or Cayley distances.*

The above model cannot be solved in reality, since the distributions of $\boldsymbol{X}^A$, $\boldsymbol{X}^B$, Y^A, and Z^A are not known. As a consequence, the following unbiased empirical estimators are used : $\widehat{\mu}_n^{\boldsymbol{X}^A}$ of $\mu_n^{\boldsymbol{X}^A}$ and $\widehat{\mu}_n^{\boldsymbol{X}^B}$ of $\mu_n^{\boldsymbol{X}^B}$. Observations of Y and Z are only available in A and B (respectively), so two empirical estimators are defined :

$$\begin{aligned} \widehat{\mu}_{n,y}^{Y^A} &= \frac{1}{n}\sum_{i\in A} \mathbb{1}_{\{Y_i=y\}},\ \forall y \in \mathcal{Y} \\ \widehat{\mu}_{n,z}^{Z^A} &= \frac{1}{n}\sum_{j\in B} \mathbb{1}_{\{Z_j=z\}},\ \forall z \in \mathcal{Z}. \end{aligned} \tag{3.8}$$

The assumption 3.2 gives that : $\mu^{Z^A} = \mu^{Z^B}$. Finally, denoting as

$$\kappa_{n,y,z} \equiv \sum_{i\in A}\sum_{j\in B} \mathbb{1}_{\{Y_i=y, Z_j=z\}}$$

the number of pairs $(i,j) \in A \times B$ such that $Y_i = y$ and $Z_j = z$, the cost matrix c is estimated by

$$\widehat{c}_{n,y,z} = \begin{cases} \frac{1}{\kappa_{n,y,z}} \sum_{i\in A}\sum_{j\in B} \mathbb{1}_{\{Y_i=y,Z_j=z\}} \times d(\boldsymbol{X}_i, \boldsymbol{X}_j), & \forall y \in \mathcal{Y}, z \in \mathcal{Z} : \kappa_{n,y,z} \neq 0, \\ 0, & \forall y \in \mathcal{Y}, z \in \mathcal{Z} : \kappa_{n,y,z} = 0. \end{cases} \tag{3.9}$$

Plugging the values observed for these estimators in (3.6) yield a linear programming model denoted as $\widehat{\mathcal{P}}_{0,n}$. A solution $\widehat{\boldsymbol{\gamma}}_{0,n}$ can then be interpreted as an estimator $\widehat{\mu}_n^{(Y^A,Z^A)}$ of the joint distribution of Y^A and Z^A, $\mu^{(Y^A,Z^A)}$. If this estimate is necessary, it is nevertheless insufficient here to provide the individual predictions on Z in A. These predictions are done in a second step using a nearest neighbor algorithm from which an estimation of $\mu^{Z^A|\boldsymbol{X}^A=\boldsymbol{x},Y^A=y}$ is deduced (see [VG7] for details).

Limits of this algorithm. This first method has two main drawbacks. First, it relies on the strong assumption that Y^A and Z^A follow the same distribution as Y^B and Z^B respectively. There are several contexts where it will not be true that Y has the same distributions in the two databases. For instance, this has already been observed when comparing North American NHANES study and the French National Health Survey. The "self-rated overall health" outcome is not distributed identically in the two databases, where the rates of functional limitations and education level are different [35]. A second limitation is that this OT model does not actually solve the recoding problem. It requires an independent post-treatment step where the predictions are computed with a nearest neighbor algorithm. In particular, this means that the choice made in the second step are not explicitly taken into account in the OT model. Moreover, the nearest neighbor algorithm has a greedy behavior where the quality of the predictions may depend on arbitrary choices in its execution.

Optimal transport of outcomes and explanatory variables within data sources

Motivation. Our aim was to build upon this previous work to develop a recoding method that requires less restrictive assumptions and directly targets the solution of the recoding problem. In particular, we consider that Y and Z may be distributed differently in A and B, but we still focus on databases where explanatory variables explain the outcomes Y and Z similarly in the two databases. This restriction remains necessary, because the only information we have to characterize the subjects is the set of common explanatory variables. In particular, if outcomes are independent from explanatory variables, recoding is doomed to failure unless additional information is provided.

In this work [VG2], we propose to search for an optimal transport between the two joint distributions of $(\boldsymbol{X}^A, Y^A)$ and $(\boldsymbol{X}^A, Z^A)$ with marginals $\mu^{(\boldsymbol{X}^A, Y^A)}$ and $\mu^{(\boldsymbol{X}^A, Z^A)}$ respectively. Under Kantorovich's formulation in a discrete setting, we will then search for

$$\boldsymbol{\gamma}^\star \in \text{argmin}_{\gamma \in \mathcal{D}} < \boldsymbol{c}, \boldsymbol{\gamma} >,$$

where $\boldsymbol{c}$ is a given cost matrix and $\mathcal{D}$ is the set of joint distributions with marginals $\mu^{(\boldsymbol{X}^A, Y^A)}$ and $\mu^{(\boldsymbol{X}^A, Z^A)}$. It is natural to see any element $\boldsymbol{\gamma} \in \mathcal{D}$ as the vector of joint probabilities $\mathbb{P}((\boldsymbol{X}^A = \boldsymbol{x}, Y^A = y), (\boldsymbol{X}^A = \boldsymbol{x}', Z^A = z))$ for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, $y \in \mathcal{Y}$

and $z \in \mathcal{Z}$. Since this probability nullifies for all $\boldsymbol{x} \neq \boldsymbol{x}'$, we will define $\boldsymbol{\gamma} \in \mathcal{D}$ as a vector of $\mathbb{R}^{|\mathcal{X}|\times|\mathcal{Y}|\times|\mathcal{Z}|}$, where $\gamma_{\boldsymbol{x},y,z}$ stands for an estimation of the joint probability $\mathbb{P}(\boldsymbol{X}^A = \boldsymbol{x}, Y^A = y, Z^A = z)$. These notations lead to the more detailed OT model

$$\mathcal{P}_1 : \begin{cases} v^\star = \min \ < \boldsymbol{c}, \boldsymbol{\gamma} > \\ \qquad \text{s.t.} \ \sum_{z\in Z} \gamma_{\boldsymbol{x},y,z} = \mu_{\boldsymbol{x},y}^{(\boldsymbol{X}^A, Y^A)}, \ \forall \boldsymbol{x} \in \mathcal{X}, \forall y \in \mathcal{Y}, \\ \qquad\quad \sum_{y\in Y} \gamma_{\boldsymbol{x},y,z} = \mu_{\boldsymbol{x},z}^{(\boldsymbol{X}^A, Z^A)}, \ \forall x \in \mathcal{X}, \forall z \in \mathcal{Z}, \\ \qquad\quad \gamma_{\boldsymbol{x},y,z} \geq 0, \ \forall \boldsymbol{x} \in \mathcal{X}, \forall y \in \mathcal{Y}, \forall z \in \mathcal{Z}. \end{cases} \tag{3.10}$$

The above model can be solved only if the marginals $\mu^{(\boldsymbol{X}^A, Y^A)}$ and $\mu^{(\boldsymbol{X}^A, Z^A)}$ are known. However, this is not the case, but we can build unbiased estimators $\widehat{\mu}_n^{(\boldsymbol{X}^A, Y^A)}$ and $\widehat{\mu}_n^{(\boldsymbol{X}^A, Z^A)}$ as in (3.3) and (3.5). The cost matrix (3.7) is used and estimated by (3.9). Formally we can write :

$$c_{\boldsymbol{x},y,z} = c_{y,z}, \ \forall \boldsymbol{x} \in \mathcal{X}, \forall y \in \mathcal{Y}, \forall z \in \mathcal{Z}. \tag{3.11}$$

Proposition 3.1. *For all $y \in \mathcal{Y}$, $z \in \mathcal{Z}$, $\widehat{c}_{n,y,z} \xrightarrow[n\to+\infty]{a.s.} c_{y,z}$, or, stated otherwise :*

$$\|\widehat{\boldsymbol{c}}_n - \boldsymbol{c}\|_\infty \xrightarrow[n\to+\infty]{a.s.} 0.$$

The resulting model is

$$\widehat{\mathcal{P}}_{1,n} : \begin{cases} \widehat{v}_n = \min \ < \widehat{\boldsymbol{c}}_n, \boldsymbol{\gamma} > \\ \qquad \text{s.t.} \ \sum_{z\in Z} \gamma_{\boldsymbol{x},y,z} = \widehat{\mu}_{n,\boldsymbol{x},y}^{(\boldsymbol{X}^A, Y^A)}, \ \forall \boldsymbol{x} \in \mathcal{X}, \forall y \in \mathcal{Y}, \\ \qquad\quad \sum_{y\in Y} \gamma_{\boldsymbol{x},y,z} = \widehat{\mu}_{n,\boldsymbol{x},z}^{(\boldsymbol{X}^A, Z^A)}, \ \forall \boldsymbol{x} \in \mathcal{X}, \forall z \in \mathcal{Z}, \\ \qquad\quad \gamma_{\boldsymbol{x},y,z} \geq 0, \ \forall \boldsymbol{x} \in \mathcal{X}, \forall y \in \mathcal{Y}, \forall z \in \mathcal{Z}. \end{cases} \tag{3.12}$$

An optimal solution, $\widehat{\boldsymbol{\gamma}}_{1,n}$, of $\widehat{\mathcal{P}}_{1,n}$ can be interpreted as an estimation of the distribution of $(\boldsymbol{X}^A, Y^A, Z^A)$. We thus deduce an estimation of the distribution of Z^A given the values of $\boldsymbol{X}^A$ and Y^A as

$$\widehat{\mu}_{n,z}^{Z^A|\boldsymbol{X}^A=\boldsymbol{x},Y^A=y} = \begin{cases} \frac{\widehat{\gamma}_{1,n,\boldsymbol{x},y,z}}{\widehat{\mu}_{n,\boldsymbol{x},y}^{(\boldsymbol{X}^A,Y^A)}}, & \forall \boldsymbol{x} \in \mathcal{X}, y \in \mathcal{Y}, z \in \mathcal{Z} : \widehat{\mu}_{n,\boldsymbol{x},y}^{(\boldsymbol{X}^A,Y^A)} \neq 0, \\ 0, & \forall \boldsymbol{x} \in \mathcal{X}, y \in \mathcal{Y}, z \in \mathcal{Z} : \widehat{\mu}_{n,\boldsymbol{x},y}^{(\boldsymbol{X}^A,Y^A)} = 0. \end{cases} \tag{3.13}$$

Remark 3.3. *If a practitioner requires a single estimated value for Z^A, we can still use the usual prediction provided by the maximum a posteriori decision rule :*

$$\widehat{z}_i^A = \text{argmax}_{z\in\mathcal{Z}} \left\{ \widehat{\mu}_{n,z}^{Z^A|\boldsymbol{X}^A=\boldsymbol{x}_i, Y^A=y_i} \right\}, \ \forall i \in A.$$

Due to the possible errors in the estimations of the terms of $\widehat{\mathcal{P}}_{1,n}$, the equality constraints are relaxed by adding slack variables in the constraints, such that they sum to zero and the ℓ_l-norm of the vector of slack variables is bounded. A regularization term is also added, considering $\left(\frac{\gamma_{\boldsymbol{x},y,z}}{\widehat{\mu}_{n,\boldsymbol{x}}^{\boldsymbol{X}^A}}\right)_{\boldsymbol{x}\in\mathcal{X},y\in\mathcal{Y},z\in\mathcal{Z}}$. Some regularity in the variations of the conditional distribution $\mu^{(Y^A,Z^A)|\boldsymbol{X}^A=\boldsymbol{x}}$ with respect to $\boldsymbol{x}$ is then expected (see [VG2] for details).

The two following works are still in progress (sections 3.2.2 and 3.2.3). I chose to introduce them because they are fully integrated into my reflections on data recoding and domain adaptation using optimal transport.

Optimal transport between the joint distribution of explanatory variables and estimated outcomes between data sources

Motivation. Optimal transport is used differently in our work on variable recoding and in the context of transfer learning and domain adaptation (section 3.2.3). Optimal transport minimizes the expected transportation cost with respect to the joint distribution of the variables in datasets A and B. In our variable recoding work, the cost function and marginal distributions are directly estimated from observations. The model corresponds to a transport across the variable categories. In domain adaptation work [27], the expectation is replaced by its empirical version, and the model corresponds to a transport across the samples (see equation (3.2)). We proposed a new formulation of the solution presented in [VG2] for statistical matching as an extension of the model developed in [27].

Model. Indeed, we extend the model developed by [27] (which will be presented in the context of domain adaptation with two variables $\boldsymbol{X}$ and Y in section 3.2.3) and propose to search for an optimal transport between the two joint distributions of $(\boldsymbol{X}^A, Y^A, Z^A)$

and $(\boldsymbol{X}^B, Y^B, Z^B)$ leading to the following OT model :

$$\gamma^\star \in \underset{\gamma \in \Gamma\left(\mu^{(\boldsymbol{X}^A, Y^A, Z^A)}, \mu^{(\boldsymbol{X}^B, Y^B, Z^B)}\right)}{\operatorname{argmin}} \sum_{((\boldsymbol{x},y,z)\in \mathcal{X}\times\mathcal{Y}\times\mathcal{Z})^2} c\Big((\boldsymbol{x}, y, z), (\boldsymbol{x}', y', z')\Big)\, \gamma_{(\boldsymbol{x},y,z),(\boldsymbol{x}',y',z')}, \tag{3.14}$$

where $c((\boldsymbol{x}, y, z), (\boldsymbol{x}', y', z')) = d(\boldsymbol{x}, \boldsymbol{x}') + \alpha_1 \mathcal{L}_{\mathcal{Y}}(y, y') + \alpha_2 \mathcal{L}_{\mathcal{Z}}(z, z')$ is a cost measure combining both the distances between the modalities and loss functions $\mathcal{L}_{\mathcal{Y}}$ and $\mathcal{L}_{\mathcal{Z}}$ measuring the discrepancy between y and y' and z and z' respectively. Since z and y' are not observed, we replace them by $f(\boldsymbol{x}, y)$ and $g(\boldsymbol{x}', z')$ respectively where $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ and $g : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ are prevision rules. We thus consider the following joint distributions $\mu^{f,A}$ and $\mu^{g,B}$ of $(\boldsymbol{X}^A, Y^A, f(\boldsymbol{X}^A, Y^A))$ and $(\boldsymbol{X}^B, g(\boldsymbol{X}^B, Z^B), Z^B)$ respectively. α_1 and α_2 are important parameters that need to be calibrated balancing the alignment of explanatory variables and outcomes. A regularization term, such as entropic, can be added, such as in [26].

The empirical version of (3.14) is given by

$$\widehat{\mathcal{P}}_{2,n} : \begin{cases} \min\limits_{f,g,\gamma} \sum\limits_{i\in A, j\in B} c\Big((\boldsymbol{x}_i, y_i, f(\boldsymbol{x}_i, y_i)), (\boldsymbol{x}_j, g(\boldsymbol{x}_j, z_j), z_j)\Big)\gamma_{i,j} \\ \text{s.t.} \sum\limits_{i\in A} \gamma_{i,j} = \dfrac{1}{n},\ \forall j \in B, \\ \quad \sum\limits_{j\in B} \gamma_{i,j} = \dfrac{1}{n},\ \forall i \in A, \\ \quad \gamma_{i,j} \geq 0,\ \forall i \in A, \forall j \in B. \end{cases} \tag{3.15}$$

The problem can be written as

$$\min_{f,g} \mathcal{W}_1\Big(\widehat{\mu}^{A,f}, \widehat{\mu}^{B,g}\Big),$$

where $\mathcal{W}_1$ is the 1-Wasserstein "distance" for the cost c and

$$\widehat{\mu}^{A,f} = \frac{1}{n}\sum_{i\in A} \delta_{(\boldsymbol{x}_i, y_i, f(\boldsymbol{x}_i, y_i))} \text{ and } \widehat{\mu}^{B,g} = \frac{1}{n}\sum_{j\in B} \delta_{(\boldsymbol{x}_j, g(\boldsymbol{x}_j, z_j), z_j)}.$$

We will assume that f and g belong to the function space $\mathcal{H}$ which is a Reproducing Kernel Hilbert Space or a function space parameterized by some parameters $\boldsymbol{w} \in \mathbb{R}^p$. For example, linear models, neural networks and kernel methods belong to such a space.

Problem $\widehat{\mathcal{P}}_{2,n}$ is smooth and the constraints are separable according to $\boldsymbol{\gamma}$, f and g. Hence, a natural way to solve the problem in equation (3.15) is to alternate optimization on parameters $\boldsymbol{\gamma}$, f and g. This algorithm known as Block Coordinate Descent (BCD) or Gauss-Seidel is described in algorithm 1. Solving $\widehat{\mathcal{P}}_{2,n}$ when f and g are fixed is a classic linear programming problem. The optimization problem with fixed $\boldsymbol{\gamma}$ and f leads to a new learning problem given by :

$$\min_{g\in\mathcal{H}} \sum_{i\in A, j\in B} \mathcal{L}_{\mathcal{Y}}\Big(g(\boldsymbol{x}_j, z_j), y_i\Big)\gamma_{i,j}. \tag{3.16}$$

The optimization problem is similar with fixed $\boldsymbol{\gamma}$ and g.

Equation (3.16) represents a multiclass classification problem. A classical approach to solve it is a one-against-all strategy. Following the example of [27], we choose this approach in our simulation study. For estimating a multiclass classifier with a one-against-all strategy, we start with a loss function $\mathcal{L} : \{0,1\} \times \{0,1\} \to \mathbb{R}$ associated to a binary classification problem and define a general loss $\mathcal{L}_{\mathcal{Y}}$, as $\mathcal{L}_{\mathcal{Y}}\Big(y_1, y_2\Big) = \sum_{y\in\mathcal{Y}} \mathcal{L}(\mathbb{1}_{\{y_1=y\}}, \mathbb{1}_{\{y_2=y\}})$. For any decision function $g \in \mathcal{H}$ and all $y \in \mathcal{Y}$, we denote as $g_y : \mathcal{X} \times \mathcal{Z} \to \{0,1\}$ the decision function related to the y-vs-all problem, $g_y(\boldsymbol{x}, z) = 1$ if $g(\boldsymbol{x}, z) = y$ and 0 otherwise. Let $\boldsymbol{P}^A$ be a $n \times |\mathcal{Y}|$ matrix such that $P^A_{i,y} = 1$ if the unit i is of class $y \in \mathcal{Y}$ and $P^A_{i,y} = 0$ otherwise. Furthermore, let $\widehat{\boldsymbol{P}}^B$ be the transported class proportion matrix given by $\widehat{\boldsymbol{P}}^B = n\boldsymbol{\gamma}^T\boldsymbol{P}^A$. Then the minimization problem given by (3.16) can be rewritten as :

$$\min_{g\in\mathcal{H}} \sum_{j\in B}\sum_{y\in\mathcal{Y}} \widehat{P}^B_{j,y}\mathcal{L}_{\mathcal{Y}}(1, g_y(\boldsymbol{x}_j, z_j)) + (1 - \widehat{P}^B_{j,y})\mathcal{L}_{\mathcal{Y}}(0, g_y(\boldsymbol{x}_j, z_j)). \tag{3.17}$$

Remark 3.4. *We can remark that solving $\widehat{\mathcal{P}}_{2,n}$ is equivalent to solve*

$$\widehat{\mathcal{P}}'_{2,n}\begin{cases} \min\limits_{f,g,\boldsymbol{p}} \sum\limits_{\boldsymbol{x}\in\mathcal{X}}\sum\limits_{y\in\mathcal{Y}}\sum\limits_{\boldsymbol{x}'\in\mathcal{X}}\sum\limits_{z\in\mathcal{Z}} c\Big((\boldsymbol{x}, y, f(\boldsymbol{x}, y)), (\boldsymbol{x}', g(\boldsymbol{x}', z), z)\Big)p_{\boldsymbol{x},y,\boldsymbol{x}',z} \\ \quad s.t. \sum\limits_{\boldsymbol{x}'\in\mathcal{X}}\sum\limits_{z\in Z} p_{\boldsymbol{x},y,\boldsymbol{x}',z} = \widehat{\mu}^{(\boldsymbol{X}^A, Y^A)}_{n,\boldsymbol{x},y},\ \forall \boldsymbol{x} \in \mathcal{X}, y \in \mathcal{Y}, \\ \quad \sum\limits_{\boldsymbol{x}\in\mathcal{X}}\sum\limits_{y\in Y} p_{\boldsymbol{x},y,\boldsymbol{x}',z} = \widehat{\mu}^{(\boldsymbol{X}^B, Z^B)}_{n,\boldsymbol{x}',z},\ \forall \boldsymbol{x}' \in \mathcal{X}, z \in \mathcal{Z}, \\ \quad p_{\boldsymbol{x},y,\boldsymbol{x}',z} \geq 0,\ \forall \boldsymbol{x} \in \mathcal{X}, y \in \mathcal{Y}, \boldsymbol{x}' \in \mathcal{X}, z \in \mathcal{Z}. \end{cases} \tag{3.18}$$

Algorithm 1: Algorithm for solving $\widehat{\mathcal{P}}_{2,n}$

Initialisation : let two classifiers $f^\star$ and $g^\star$;
while *do not converge* **do**
  Solve $\boldsymbol{\gamma}^\star \in$
  $$\underset{\boldsymbol{\gamma}\in\Gamma\left(\frac{1}{n}\sum_{i\in A}\delta_i,\frac{1}{n}\sum_{j\in B}\delta_j\right)}{\operatorname{argmin}} \sum_{i\in A,j\in B} c\Big((\boldsymbol{x}_i,y_i,f^\star(\boldsymbol{x}_i,y_i)),(\boldsymbol{x}_j,g^\star(\boldsymbol{x}_j,z_j),z_j)\Big)\,\gamma_{i,j}\ ;$$
  $\boldsymbol{\gamma}^\star = \mathrm{OT}(\boldsymbol{w},\boldsymbol{w},\boldsymbol{C})$ with $\boldsymbol{w}=\mathbf{1}_n$ the vector of n-ones and $C_{i,j} = c\Big((\boldsymbol{x}_i,y_i,f^\star(\boldsymbol{x}_i,y_i)),(\boldsymbol{x}_j,g^\star(\boldsymbol{x}_j,z_j),z_j)$;
  Solve $f^\star \in \operatorname{argmin}_{f\in\mathcal{H}} \sum_{i\in A,j\in B}\mathcal{L}\Big(f(\boldsymbol{x}_i,z_i),y_j\Big)\,\gamma^\star_{i,j}$;
  Solve $g^\star \in \operatorname{argmin}_{g\in\mathcal{H}} \sum_{i\in A,j\in B}\mathcal{L}\Big(z_i,g(\boldsymbol{x}_j,y_j)\Big)\,\gamma^\star_{i,j}$;
Return $\widehat{z}_i = f^\star(\boldsymbol{x}_i,y_i),\forall i\in A$ and $\widehat{y}_j = g^\star(\boldsymbol{x}_j,y_j),\forall j\in B$;

Theoretical properties

Consistency of the optimal transport estimator for model defined in section 3.2.2. In [VG2], we study the asymptotic behavior of a optimal solutions of $\widehat{\mathcal{P}}_{1,n}$ with respect to those of $\mathcal{P}_1$ (see equation (3.10)), the models defined in section 3.2.2. To do this we rewrite $\mathcal{P}_1$ and $\widehat{\mathcal{P}}_{1,n}$ as generic linear programs in standard form

$$\mathcal{P}_1 : v^\star = \min\left\{\langle \boldsymbol{c}|\boldsymbol{\gamma}\rangle : \boldsymbol{A}\boldsymbol{\gamma}=\boldsymbol{b},\boldsymbol{\gamma}\geq\mathbf{0}\right\} \quad \text{and} \quad \widehat{\mathcal{P}}_{1,n} : \widehat{v}_n = \min\left\{\langle \widehat{\boldsymbol{c}}_n|\boldsymbol{\gamma}\rangle : \boldsymbol{A}\boldsymbol{\gamma}=\widehat{\boldsymbol{b}}_n,\boldsymbol{\gamma}\geq\mathbf{0}\right\},$$

where $\boldsymbol{c},\widehat{\boldsymbol{c}}_n\in\mathbb{R}^p$, $\boldsymbol{b},\widehat{\boldsymbol{b}}_n\in\mathbb{R}^m$, $\boldsymbol{A}\in\mathbb{R}^{m\times p}$ and $\mathbf{0}=\mathbf{0}_p$ the vector of p-zeros. For $\boldsymbol{\gamma}\in\mathbb{R}^p$ and $S\subset\mathbb{R}^p$, we also define the point-to-set distance as

$$d(\boldsymbol{\gamma},S) = \inf_{\boldsymbol{\gamma}'\in S}\|\boldsymbol{\gamma}-\boldsymbol{\gamma}'\|,$$

where $\|\cdot\|$ is some norm on $\mathbb{R}^p$. We then introduce the deviation measure of set $S\subset\mathbb{R}^p$ from set $S'\subset\mathbb{R}^p$ as

$$\mathbb{D}(S,S') = \sup_{\boldsymbol{\gamma}\in S}\inf_{\boldsymbol{\gamma}'\in S'}\|\boldsymbol{\gamma}-\boldsymbol{\gamma}'\|.$$

Theorem 3.1. *Let $S^\star$ be the set of optimal solutions of $\mathcal{P}_1$ and $\widehat{S}_n$ the set of optimal solutions of $\widehat{\mathcal{P}}_{1,n}$. Then*

$$\widehat{v}_n \xrightarrow[n\to+\infty]{a.s.} v^\star, \qquad \textit{and} \qquad \mathbb{D}(\widehat{S}_n,S^\star) \xrightarrow[n\to+\infty]{a.s.} 0.$$

The convergence of $\mathbb{D}(\widehat{S}_n, S^\star)$ justifies that we estimate a solution of $\mathcal{P}_1$ with one of $\widehat{\mathcal{P}}_{1,n}$. However, it cannot justify the overall approach that consists in searching for a solution of $\mathcal{P}_1$ to derive the conditional distributions $\mu^{Z^A|\boldsymbol{X}^A=x,Y^A=Y}, \boldsymbol{x} \in \mathcal{X}, y \in \mathcal{Y}$. The quality of our estimation of these distributions will depend on how the cost function reflects some unobserved properties of the distributions. In the choice of our cost function, we follow the intuition that subjects with outcomes $Y_i = y$ and $Z_i = z$ should be frequent when y and z are close to each other in the space of the explanatory variables. If for instance, there is some prior distribution $\mu^{(\boldsymbol{X}^A,Y^A,Z^A)}$, the result would certainly be improved by minimizing $\left\|\boldsymbol{\gamma} - \mu^{(\boldsymbol{X}^A,Y^A,Z^A)}\right\|_1$ instead.

Bound on the prediction error for models defined in sections 3.2.2 and 3.2.2. Let $(\mathcal{X}, d_{\mathcal{X}})$, $(\mathcal{Y}, \mathcal{L}_{\mathcal{Y}})$ and $(\mathcal{Z}, \mathcal{L}_{\mathcal{Z}})$ metric spaces. Let $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ be a prevision rule. We define the risk or error of generalization of f by

$$\mathcal{R}^j(f) = \mathbb{E}_{(\boldsymbol{X},Y,Z)\sim\mu^{(\boldsymbol{X}^j,Y^j,Z^j)}}[\mathcal{L}_{\mathcal{Z}}(Z, f(\boldsymbol{X}, Y))]$$

where $j = A, B$. Let define, $\forall \boldsymbol{x} \in \mathcal{X}, y \in \mathcal{Y}$,

$$\begin{aligned} f_1^\star(x,y) &\in \operatorname{argmin}_{z'\in\mathcal{Z}} \mathbb{E}[\mathcal{L}_{\mathcal{Z}}(Z^A, z')|\boldsymbol{X}^A = \boldsymbol{x}, Y^A = y] \\ &= \operatorname{argmin}_{z'\in\mathcal{Z}} \sum_{z\in\mathcal{Z}} \mathcal{L}_{\mathcal{Z}}(z, z')(\mu^\star)_z^{Z^A|\boldsymbol{X}^A=\boldsymbol{x},Y^A=y} \qquad (3.19) \\ \text{and } \widehat{f}_{1,n}(x,y) &\in \operatorname{argmin}_{z'\in\mathcal{Z}} \sum_{z\in\mathcal{Z}} \mathcal{L}_{\mathcal{Z}}(z, z')\widehat{\mu}_{n,z}^{Z^A|\boldsymbol{X}^A=\boldsymbol{x},Y^A=y} \qquad (3.20) \end{aligned}$$

with $\widehat{\mu}_{n,z}^{Z^A|\boldsymbol{X}^A=\boldsymbol{x},Y^A=y}$ defined in equation (3.13) from the model $\widehat{\mathcal{P}}_{1,n}$, and $(\mu^\star)_z^{Z^A|\boldsymbol{X}^A=\boldsymbol{x},Y^A=y}$ defined in similar way, by replacing $\widehat{\boldsymbol{\gamma}}_n$ with $\boldsymbol{\gamma}^\star$, a solution of $\mathcal{P}_1$ and $\widehat{\mu}_n^{(\boldsymbol{X}^A,Y^A)}$ with $\mu^{(\boldsymbol{X}^A,Y^A)}$ (section 3.2.2). We denote by $\widehat{f}_{2,n}$ a solution of $\widehat{\mathbb{P}}_{2,n}$, the model defined in section 3.2.2.

For both prediction functions $\widehat{f}_{1,n}$ and $\widehat{f}_{2,n}$, we proposed bounds on the error of generalization in our current work.

Conclusion

Proposed methods and simulation results. We have proposed several methods for predicting missing variables Y and Z using the theory of optimal transport.

- In our first work, the distribution of Y is transported to the distribution of Z within a data source, and a cost function is proposed as an average distance between the covariate vector $\boldsymbol{X}$ of each data source [VG7]. This approach has shown better performances when compared to missing data imputation methods such as multiple imputation [109], non-parametric hot-deck approaches [42], or a statistical learning method. In this work, we assume that the distributions of Y, Z, $Y|\boldsymbol{X}$, and $Z|\boldsymbol{X}$ remain unchanged from one data source to another.
- This algorithm is extended in [VG2], considering the transport of the joint distribution between $(\boldsymbol{X}, Y)$ and $(\boldsymbol{X}, Z)$ within a data source. The constraints on the marginals of OT problem are also relaxed, because they may be too restrictive in the presence of errors in the estimations. A regularization term is added to the objective function to smooth the variations of outcomes with respect to explanatory variables. Only the distributions of $Y|\boldsymbol{X}$ and $Z|\boldsymbol{X}$ are assumed to remain unchanged across data sources. This method has demonstrated better performances than the previous method in most different scenarios under different assumptions on the data.
- We proposed a new formulation for data recoding as an extension of the model developed in [27]. Two classifiers, f and g, are introduced to predict missing values of Z in A and Y in B. The proposed algorithm minimizes the optimal transport loss between the joint distribution of $(\boldsymbol{X}, Y, g(\boldsymbol{X}, Y))$ in database A and the joint distribution of $(\boldsymbol{X}, f(\boldsymbol{X}, Y), Y)$ in database B. Now, we are comparing this method with the approach proposed in [VG2] under different assumptions on the data distributions. Simulations are still in progress. For both algorithms, we provided a bound on the prediction error. This new formulation has the advantage of handling cases where Y and Z are continuous variables, which was not possible with previous algorithms.

3.2.3 Heterogeneous domain adaptation

Notations and problem setting

In the following, the two databases are denoted by $S = \{i_1, \ldots, i_{n_S}\}$ and $T = \{j_1, \ldots, j_{n_T}\}$. $((\boldsymbol{X}_i, Y_i))_{i \in S}$ and $((\boldsymbol{X}_j, Y_j))_{j \in T}$ be two sequences of i.i.d. discrete random variables with values in $\mathbb{R}^{d^S} \times \mathcal{Y}$ and $\mathbb{R}^{d^T} \times \mathcal{Y}$, respectively, where $\mathcal{Y}$ is $\mathbb{R}$ or a finite subset of $\mathbb{R}$. Variables $(\boldsymbol{X}_i, Y_i), i \in S$ are i.i.d copies of $(\boldsymbol{X}^S, Y^S)$ and $(\boldsymbol{X}_j, Y_j), j \in T$,

are i.i.d copies of $(\boldsymbol{X}^T, Y^T)$. $d(.)$ denotes a distance relative to the problem. The different objectives are presented in section 3.2.1.

Background on Optimal Transport for Domain Adaptation

Optimal Transport for Homogeneous Domain Adaptation. Here, explanatory variables are in the same space and $d^S = d^T$. OT can solve a DA problem (OTDA, [28]) assuming that $\mu^{Y^T|\boldsymbol{X}^T} = \mu^{Y^S|f(\boldsymbol{X}^S)}$ and $\mu^{\boldsymbol{X}^T} = \mu^{f(\boldsymbol{X}^S)}$ with $f : \mathbb{R}^{d^S} \mapsto \mathcal{Y}$ a nonlinear transformation of the input space, the transport map $\boldsymbol{\pi}$ is obtained by solving the OT problem (defined in equation (3.2)) between $\boldsymbol{X}^S$ and $\boldsymbol{X}^T$: $\boldsymbol{\pi} = \text{OT}(\boldsymbol{w}^S, \boldsymbol{w}^T, \boldsymbol{C})$ where $\boldsymbol{w}^S = \mathbf{1}_{n^S}/n^S$, $\boldsymbol{w}^T = \mathbf{1}_{n^T}/n^T$ and $\boldsymbol{C} \in \mathbb{R}^{n^S \times n^T}$ such that $C_{i,j} = d(\boldsymbol{x}_i^S, \boldsymbol{x}_j^T)$. The transport map between samples $\boldsymbol{\pi} \in \mathbb{R}^{n^S \times n^T}$ is then used to calculate the barycentric coordinates of the source samples in the target domain : $M(\boldsymbol{x}_i^S) = \frac{1}{w_i^S} \sum_{j \in T} \pi_{i,j} \boldsymbol{x}_j^T$. Finally, a predictive function f^T is trained on the transported source data $((M(\boldsymbol{x}_i^S), y_i^S))_{i \in S}$ and applied on the target data $\boldsymbol{x}_j^T$, $j \in T$ to predict target outcomes : $y_j^T = f^T(\boldsymbol{x}_j^T)$, $j \in T$. Developed to deal with changes in the marginal distributions of the features and in the conditional distributions of Y that can occur with real-world data but that are not handled by OTDA, Joint Distribution Optimal Transport (JDOT) [27] consists in simultaneously optimizing the coupling matrix $\boldsymbol{\pi}$ and the predictive function f^T by minimizing :

$$\min_{\boldsymbol{\pi}, f^T} \sum_{i \in S} \sum_{j \in T} \Big[\alpha \; d(\boldsymbol{x}_i^S, \boldsymbol{x}_j^T) + \mathcal{L}(y_i^S, f^T(\boldsymbol{x}_j^T)) \Big] \pi_{i,j} \tag{3.21}$$

where α is a hyper-parameter and $\mathcal{L}$ is a loss function. A Block Coordinate Descent (BCD) or Gauss-Seidel is performed to alternatively estimate $\boldsymbol{\pi}$ and f^T in practice. The authors have shown the superiority of their approach through experiments on benchmark datasets *w.r.t.* several DA state-of-the-art methods, including previous OT-based approaches, domain adversarial neural networks, or transfer components. A deep-learning version has also been proposed [30].

Optimal Transport for Heterogeneous Domain Adaptation. The way OT maps two domains as in OTDA or JDOT makes it impractical if these domains are in different spaces. Co-Optimal Transport (COOT) [93] has therefore been developed to deal with incomparable spaces, by simultaneously optimizing two transport maps between both samples ($\boldsymbol{\pi}^{\boldsymbol{s}}$) and variables ($\boldsymbol{\pi}^{\boldsymbol{v}}$), thus allowing comparison between normally incomparable distributions while providing two usable transport plans. The COOT problem is

defined as :

$$\min_{\substack{\boldsymbol{\pi}^s \in \Pi(\boldsymbol{w}^S, \boldsymbol{w}^T) \\ \boldsymbol{\pi}^v \in \Pi(\boldsymbol{v}^S, \boldsymbol{v}^T)}} \sum_{i \in S} \sum_{j \in T} \sum_{k=1}^{d_S} \sum_{\ell=1}^{d_T} \mathcal{L}(x_{i,k}^S, x_{j,\ell}^T) \pi_{i,j}^s \pi_{k,\ell}^v \tag{3.22}$$

where $\boldsymbol{v}^S$ and $\boldsymbol{v}^T$ are the weights associated with the variables $X_1^S, \ldots, X_{d_S}^S$ and $X_1^T, \ldots, X_{d_T}^T$ respectively. Both transport plans are optimized through a BCD, given the natural separability of the problem, and is an improvement over other similar methods, in particular [119] who consider a semi-supervised entropic Gromov-Wasserstein discrepancy.

Formulation of the algorithms. Let $(\boldsymbol{X}^S, \boldsymbol{X}^T) \in \mathbb{R}^{d^S + d^T}$ a vector of random variables and $(K, L) \in \left\{1, \ldots, d^S\right\} \times \left\{1, \ldots, d^T\right\}$ with K the random variable corresponding to the index K of the Kth variable of the vector $\boldsymbol{X}^S$ and L the random variable corresponding to the index L of the Lth variable of the vector $\boldsymbol{X}^T$. We assume that $(\boldsymbol{X}^S, \boldsymbol{X}^T)$ and (K, L) are independent. The distribution of K corresponds to the weight vector $\boldsymbol{v}^S$ and the distribution of L to $\boldsymbol{v}^T$. Let

$$\begin{aligned} P^S &: \mathbb{R}^{d^S} \times \left\{1, \ldots, d^S\right\} \to \mathbb{R} \\ &\quad (x_1^S, \ldots, x_{d^S}^S, k) \mapsto x_k^S, \\ \text{and } P^T &: \mathbb{R}^{d^T} \times \left\{1, \ldots, d^T\right\} \to \mathbb{R} \\ &\quad (x_1^T, \ldots, x_{d^T}^T, \ell) \mapsto x_\ell^T. \end{aligned}$$

The OTDA algorithm is an empirical vesion of the model which looks for $\pi^S \in \Gamma\left(\mu^{\boldsymbol{X}^S}, \mu^{\boldsymbol{X}^T}\right)$ that minimises $\mathbb{E}[c(\boldsymbol{X}^S, \boldsymbol{X}^T)]$. The COOT algorithm is an empirical vesion of the model which looks for $\pi^S \in \Gamma\left(\mu^{\boldsymbol{X}^S}, \mu^{\boldsymbol{X}^T}\right)$ and $\boldsymbol{\pi}^v \in \Gamma\left(\mu^K, \mu^L\right) = \Pi(\boldsymbol{v}^S, \boldsymbol{v}^T)$ that minimises $\mathbb{E}[c(\boldsymbol{X}_K^S, \boldsymbol{X}_L^T)]$. We then have :

$$\begin{aligned} \mathbb{E}[c(\boldsymbol{X}_K^S, \boldsymbol{X}_L^T)] &= \mathbb{E}[c(P^S(\boldsymbol{X}^S, K), P^T(\boldsymbol{X}^T, L))] \\ &= \sum_{k,\ell} \int_{\mathbb{R}^{d^S \times d^T}} c(x_k^S, x_\ell^T) \pi^s(d\boldsymbol{x}^S, d\boldsymbol{x}^T) \pi_{k,\ell}^v \end{aligned}$$

because $(\boldsymbol{X}^S, \boldsymbol{X}^T)$ and (K, L) are independent.

Optimal Transport for Heterogeneous Domain Adaptation

The aim of the proposed method called Joint Distribution Co-Optimal Transport (JDCOOT) is to deal with *heterogeneous* domain adaptation, with different feature spaces and distribution shift across domains. We propose to use the principle of COOT, which allows to match samples and features from incomparable spaces, to adapt JDOT to the HDA framework (see figure 3.5).

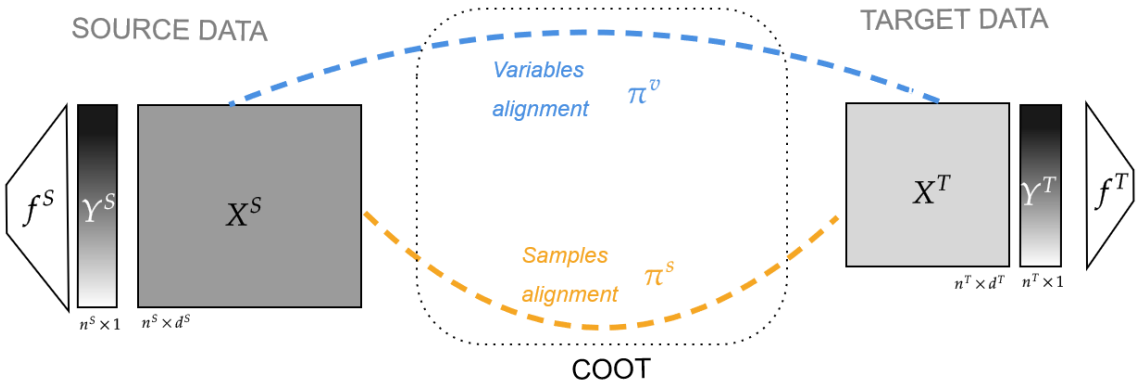

FIGURE 3.5 – Heterogenous Domain Adaptation with Optimal Transport : JDCOOT simultaneously solves the OT problem on the samples ($\boldsymbol{\pi^s}$) and variables ($\boldsymbol{\pi^v}$) as in COOT, and the classification problems (f^S and f^T) as in JDOT.

Method. Let's recall that the set of the sample indices where the outcome Y is observed in the source and target samples are denoted respectively by $\mathcal{I}^S$ and $\mathcal{I}^T$. The JDCOOT method consists in simultaneously solving the OT problem on the samples ($\boldsymbol{\pi^s}$), the OT problem on the variables ($\boldsymbol{\pi^v}$) and the classification problems (f^S and f^T) :

$$\min_{\substack{\boldsymbol{\pi^s},\boldsymbol{\pi^v},f^S,f^T \\ \boldsymbol{\pi^s}\in\Pi(\boldsymbol{w}^S,\boldsymbol{w}^T) \\ \boldsymbol{\pi^v}\in\Pi(\boldsymbol{v}^S,\boldsymbol{v}^T)}} \sum_{i\in S}\sum_{j\in T}\sum_{k=1}^{d_S}\sum_{l=1}^{d_T}\left(\alpha\, d(x^S_{i,k},x^T_{j,\ell}) + \mathcal{L}(\widetilde{f}^S(\mathbf{x}^S_i),\widetilde{f}^T(\mathbf{x}^T_j))\right)\pi^s_{i,j}\pi^v_{k,\ell}, \tag{3.23}$$

where α is a hyper-parameter. $\widetilde{f}^S(\boldsymbol{x}_i) = y_i$ if $i \in \mathcal{I}^S$ (i.e. y_i is observed) and $\widetilde{f}^S(\boldsymbol{x}_i) = f^S(\boldsymbol{x}_i)$, otherwise. $\widetilde{f}^T$ is defined similarly on the target domain.

Optimization. Optimization is done by minimizing (3.23) with a BCD, as described in Algo. 2. $\boldsymbol{\pi^s}$ and $\boldsymbol{\pi^v}$ are initialized randomly. f^S and f^T are initialized with available

data with observed outcomes if any. In the unsupervised case, f^T is initialized considering transported source data by the transport map given by the algorithm COOT.

Algorithm 2: BCD for JDCOOT

Initialisation : $\boldsymbol{\pi}^{\boldsymbol{s}}_{(0)} \leftarrow \boldsymbol{\pi}^{\boldsymbol{s}}_{init}$; $\boldsymbol{\pi}^{\boldsymbol{v}}_{(0)} \leftarrow \boldsymbol{\pi}^{\boldsymbol{v}}_{init}$; $\widehat{f}^S_{(0)} \leftarrow \widehat{f}^S_{init}$; $\widehat{f}^T_{(0)} \leftarrow \widehat{f}^T_{init}$

for $k = 1 \ldots K$ **do**

  $\boldsymbol{\pi}^{\boldsymbol{s}}_{(k+1)} \leftarrow \mathrm{OT}(\boldsymbol{w}^S, \boldsymbol{w}^T, \boldsymbol{C} \otimes \boldsymbol{\pi}^{\boldsymbol{v}}_{(k)})$ where $\otimes$ is the Kronecker product

  $\boldsymbol{\pi}^{\boldsymbol{v}}_{(k+1)} \leftarrow \mathrm{OT}(\boldsymbol{v}^S, \boldsymbol{v}^T, \boldsymbol{C} \otimes \boldsymbol{\pi}^{\boldsymbol{s}}_{(k)})$

  Train classifiers (gradient descent) $\widehat{f}^S_{(k)}$ to minimize (3.23) using, $\widehat{f}^T_{(k)}$ $\boldsymbol{\pi}^{\boldsymbol{s}}_{(k+1)}$ and $\boldsymbol{\pi}^{\boldsymbol{v}}_{(k+1)}$

  Train classifiers (gradient descent) $\widehat{f}^T_{(k)}$ to minimize (3.23) using, $\widehat{f}^S_{(k+1)}$ $\boldsymbol{\pi}^{\boldsymbol{s}}_{(k+1)}$ and $\boldsymbol{\pi}^{\boldsymbol{v}}_{(k+1)}$

Simulation scenarios

I introduced scenarios designed to represent a particular case of domain heterogeneity which also mimics the structure of our real data example.

The simulation study is based on several scenarios for data characteristics that are controlled to evaluate their impact on JDCOOT performances. In our numerical study, an underling relationship between the outcome and some explanatory variables (called *active variables*) is assumed to be the same in source and target domains, heterogeneity for domain adaptation coming from a masking strategy to define the set of actually (active and non active) observed variables which differs between the 2 domains. Therefore, we consider the following 2-step data generation process : (step1) generate the explanatory variables $\boldsymbol{X}$ and the outcome variables, denoted Y for a regression problem and Z for a binary classification (logistic) problem and (step2) apply a mask strategy for handling observed variables and imitate the heterogeneous case beteween the 2 domains.

Step 1 : Generation of explanatory variables $\boldsymbol{X}$, and response variables Y and Z. n independent samples of $(\boldsymbol{X}, Y, Z)$ are generated according to predefined distributions that may differ in source S and target T.

Explanatory variables. Let $\boldsymbol{X} = (X_1, \ldots, X_d)$ be a family of *i.i.d.* d-dimensional random vectors with multivariate normal distribution $\mathcal{N}(\boldsymbol{m}, \boldsymbol{\Sigma})$.

Outcome variables. Two response variables are then generated, considering regression and classification problems. In each case, the coefficients $\mathbf{a} = (a_1, \ldots, a_d)$ associated to the d variables are such that p coefficients are fixed to $a_k \neq 0$ (indexed by $k_1, \ldots, k_p$) and the $d - p$ others (indexed by $k_{p+1}, \ldots, k_d$) are fixed to $a_k = 0$. When $a_k \neq 0$, the associated variable is called *active variable*, as it corresponds to a significant link with the response variables. We call *sparse rate* the proportion of active variables : $\text{sr} = \frac{p}{d}$. From this, we define $\mathbf{\Sigma}$ be a 2×2-block matrix. The diagonal blocks take values $\Sigma_{k_\ell k_{\ell'}} = \text{Cor}(X_{k_\ell}, X_{k_{\ell'}}) = \rho_1^{|\ell-\ell'|}, \forall \ell, \ell' \in [1; p]$ and $\Sigma_{k_\ell k_{\ell'}} = \text{Cor}(X_{k_\ell}, X_{k_{\ell'}}) = \rho_2^{|\ell-\ell'|}, \forall \ell, \ell' \in [p+1; d]$ respectively.

- Regression : Y is a continuous variable that depends linearly on $\mathbf{X}$ as follows :

$$Y = \mathbf{a}^\top \mathbf{X} + \epsilon, \tag{3.24}$$

  with $\epsilon \sim \mathcal{N}(0, \sigma^2)$. σ is set so that the coefficient of determination R^2 of Y from $\boldsymbol{X}$ reaches a given value, with :

$$R^2 = \frac{\mathbb{V}(\mathbf{a}^\top X)}{\mathbb{V}(\mathbf{a}^\top \boldsymbol{X}) + \sigma^2} = \frac{\sum_{k_\ell,\ell=1}^{p} \mathbb{V}(\boldsymbol{X}_{k_\ell}) + 2\sum_{\ell=1}^{p}\sum_{\ell'=1,k_\ell<k_{\ell'}}^{p} \text{Cor}(\boldsymbol{X}_{k_\ell}, \boldsymbol{X}_{k_{\ell'}})}{\sum_{k_\ell,\ell=1}^{p} \mathbb{V}(\boldsymbol{X}_{k_\ell}) + 2\sum_{\ell=1}^{p}\sum_{\ell'=1,k_\ell<k_{\ell'}}^{p} \text{Cor}(\boldsymbol{X}_{k_\ell}, \boldsymbol{X}_{k_{\ell'}}) + \sigma^2}.$$

- Classification : Z is a binary variable such that, for $z = 0, 1$

$$\mathbb{P}(Z = z|\boldsymbol{X} = \boldsymbol{x}) = \frac{e^{\boldsymbol{a}^\top \boldsymbol{x}}}{1 + e^{\mathbf{a}^\top \boldsymbol{x}}} \tag{3.25}$$

  a_k for active variables are set with $\text{OR}_k(x) = \text{OR}(x) = \exp(a_k)$ [1], $\forall k \in \{k_1, \ldots, k_p\}$ the odds ratio reaches a given value.

Step2 : Mask strategy for handling observed data. In source and target data, the underlying relation between response and explanatory variables are the same. To simulate an heterogeneous domain case, we consider two masks for observed explanatory variables. Each mask allows to control the number of observed variables in each domain (d^S and d^T) such as the proportion of active observed variables in each domain is fixed (p_x^S and

1.
$$\text{OR}_k = \frac{\mathbb{P}(Z = 1|x_k = x \text{ and } x_{k'} = 0, \forall k' \neq k)/\mathbb{P}(Z = 0|x_k = x \text{ and } x_{k'} = 0, \forall k' \neq k)}{\mathbb{P}(Z = 1|x_k = x + 1 \text{ and } x_{k'} = 0, \forall k' \neq k)/\mathbb{P}(Z = 0|x_k = x + 1 \text{ and } x_{k'} = 0, \forall k' \neq k)}$$

p_x^T) among the p variables with $a_k \neq 0$. According to the learning case, the outcomes are observed on a proportion p_o of samples in source and/or target data.

Theoretical properties

We proposed bounds on the error of generalization of the prediction function $\widehat{f}_n^S$ and $\widehat{f}_n^T$ solutions of the problem defined in equation (3.23) in our current work.

Conclusion

Proposed method. In our work, we focused on the utilization of CO-Optimal Transport for Heterogeneous Domain Adaptation (HDA). A classifier f is introduced to predict a categorical or continuous target variable Y given an input $\boldsymbol{X} = \boldsymbol{x}$. The proposed algorithm minimizes the loss function of CO-Optimal Transport between the source joint distribution $(\boldsymbol{X}, Y)$ and the estimated target joint distribution $(\boldsymbol{X}, f(\boldsymbol{X}))$. We have also proposed a bound on the prediction error.

Simulation results. Simulations are still in progress. This work will be applied to real omic data.

3.3 Record linkage : extending the Fellegi-Sunter record linkage model

Motivations :

- **Probabilistic record linkage** (chaining) : a process of combining data from different sources when these data refer to common entities and identification information is not available.
- We can only use other partial identifiers (called matching variables) which are common to both databases to identify matched pairs from the two databases. Our data include various types of matching variables, including binary, categorical, and continuous variables.

Contributions of the Chapter :

- Extending the Fellegi-Sunter record linkage model for mixed-type comparison values [VG14].
  - ✓ Two novel comparison approaches for low prevalence categorical matching variables and continuous matching variables.

Collaborations :

- PhD of *Huan Tanh Vo*. Co-direction with *Guillaume Chauvet* (ENSAI, Rennes), *Andre Happe* (Université de Rennes 1) and *Stephane Paquelet* (IRT bcom).

Perspectives :

- Record linkage using optimal transportation.
- Package R.

3.3.1 Introduction

Problem and objective

Problem. The first part of *Huan Vo Tanh*'s PhD work concerns probabilistic record linkage (chaining), which is a process of combining data from different sources when these data refer to common entities and identification information is not available. For example, the GETBO project is a registry of venous thromboembolism (VTE) cases between 2013 and 2015 in Brest, France. The registry only recorded the demographic information for each patient (date of birth, gender, residency code) along with some dates and types of medical acts. This information is not sufficient to build an analysis model such as a

prediction model which can early identify symptomatic VTE [89]. The French National Health Data System is the national health data system which collects all the longitudinal health records and insurance information of most of the French population [8]. The SNDS comprises data from health insurance, hospitalizations, causes of death, and pathologies, as well as a sample from supplementary health insurance systems. The valuable data in SNDS can be used to enrich the registry, which is expected to lead to valuable knowledge for researchers. This motivates us to link the GETBO and SNDS so that we can get more medical information for GETBO patients and thus, to improve the subsequent statistical analysis. Deterministic approaches can be possible when the combination of covariates from different individuals produces a unique patient identifier. When no unique patient identifier is available, or when such an identifier cannot be used for privacy reasons, alternative approaches must be employed. In such situation, we can only use other partial identifiers which are common to both databases (e.g., gender, postal code, or dates of medical treatments) to identify matched pairs from the two databases. These variables are often referred to as matching variables (see figure 3.6). atching variables) which are common to both databases to identify matched pairs from the two databases. Our data include various types of matching variables, including binary, categorical, and continuous variables.

Objective. The objective is to link de-identified research datasets at the patient level when a common individual identifier is not available using mixed-type matching variables.

Literature

The Fellegi-Sunter probabilistic record linkage model [47] laid the foundation for most record linkage models until now [21]. However, the Fellegi-Sunter method employs only binary comparison between matching variables. Firstly, the simple binary comparison is not sufficient to account for the characteristics of low prevalence matching variables such as cancer diagnosis codes [58]. For example, two patients who both have lung cancer are more likely to match than two patients without cancer. However, the binary comparison leads to the same matching probability for the two cases. Such cases are considered in [58], who propose a Bayesian linkage framework outperforming the Fellegi-Sunter model. However, their model is restricted to binary matching variables only. Secondly, this binary comparison is not able to account for tolerances in the continuous matching variables, such as the delay in the date of treatments reported in different sectors due to administrative

process (see figure 3.6). These are the common types of matching variables in health data as well as in GETBO and SNDS. Some authors introduced continuous similarity measures for comparing string data, but then comparison values are transferred to categorical values representing different levels of agreement [e.g., 60, 99, 46], which may result in a loss of information.

		Matching variables		
	NIR	Postal code	Echodoppler date	Cancer
a_1	~~1 96 09 99~~	29001	10/03/2014	1
a_2	~~2 93 08 88~~	29002	17/05/2013	0
a_3	~~2 85 06 66~~	29003	19/11/2013	0
a_4	~~1 99 05 55~~	29002	01/03/2014	0
⋮		⋮	⋮	⋮
a_n	~~1 86 03 65~~	29010	6/09/2016	1

SNDS

		Matching variables		
	NIR	Postal code	Echodoppler date	Cancer
b_1	~~1 96 09 99~~	29001	12/03/2014	1
b_2	~~2 93 08 88~~	29002	17/05/2013	0
⋮		⋮	⋮	⋮
b_m	~~1 90 01 86~~	29010	12/01/2015	0

GETBO registry

FIGURE 3.6 – Matching variable example.

3.3.2 Probabilistic record linkage

Consider two databases A and B containing n_A and n_B records respectively, and with elements in common. Following the terminology in [47], each possible pair of individuals (a_i, b_j) with $a_i \in A, i = 1, \ldots, n_A$ and $b_j \in B, j = 1, \ldots, n_B$ either belongs to the set of true matched pairs

$$M = \{(a, b); a = b, a \in A, b \in B\},$$

or to the set of true unmatched pairs

$$U = \{(a, b); a \neq b, a \in A, b \in B\}.$$

Because an identifying variable is not available, other less discriminant data are used in the probabilistic record linkage procedure, such as the name, date of birth, postal code, or some diagnosis codes. This information needs to be registered in both data sets and is referred to as matching variables. The matching variables in the two databases are required to have the same format [20].

It is supposed that there is no prior knowledge of how likely the matches are, which is often the case in practice. The strategy therefore begins by comparing K matching variables for all records $\mathbf{X}_{A,i} = (X^1_{A,i}, \ldots, X^K_{A,i}), i = 1, \ldots, n_A$ of n_A individuals in A, with all records $\mathbf{X}_{B,j} = (X^1_{B,j}, \ldots, X^K_{B,j}), j = 1, \ldots, n_B$ of n_B individuals in B. This leads to $n_A \times n_B$ comparison vectors $\mathbf{\Gamma}_{i,j}$ such that

$$\mathbf{\Gamma}_{i,j} = \left\{\Gamma^1_{i,j}, \ldots, \Gamma^k_{i,j}, \ldots, \Gamma^K_{i,j}\right\}, \tag{3.26}$$

where $\Gamma^k_{i,j} = h^k(X^k_{A,i}, X^k_{B,j})$ and h^k is a comparison function for the $k-$th matching variable.

Because the number of all record pairs is quadratic in the number of individuals in each database, making the comparison for all possible record pairs is often impracticable in applications. One of the most popular methods to reduce the number of record pairs that need to be compared is blocking, in which only records from the two databases that are in the same block (i.e., sharing the same values for the blocking variables) are compared with each other. Record pairs disagreeing on the blocking variable are automatically classified as non-matches. Therefore, blocking is a trade-off between computational cost and the proportion of missed matches (matched pairs are missed because of errors in the blocking variable), see [60].

The set of all possible realizations of $\mathbf{\Gamma}$ is called the comparison space $\mathcal{L}$. The comparison function h^k for the k-th matching variable can be defined in different ways depending on the type of matching variable [20]. The most common way consists of a binary comparison, i.e.

$$\Gamma^k_{i,j} = h^k(X^k_{A,i}, X^k_{B,j}) = \begin{cases} 1 & \text{if} \quad X^k_{A,i} = X^k_{B,j}, \\ 0 & \text{if} \quad X^k_{A,i} \neq X^k_{B,j}. \end{cases} \tag{3.27}$$

If there is no error in the matching data, all components of a comparison vector of a matched pair are equal to 1. However, application data usually contain errors (e.g., typographical), and some similarity measures that can take them into account have been developed in the literature for string variables [60].

Once all candidate pairs are compared, various approaches are possible to classify the set of comparison vectors into matches and non-matches [20]. If training data where we observe the true matched status of record pairs is available, supervised classification methods [19] can be used to find a classification rule. If there is no training data but some clerical review is possible, some semi-supervised approaches [e.g. 45] may be applied. However, the exact knowledge of matches is rarely possible in real-world situations, and the clerical review is costly. Unsupervised methods [e.g. 116, 83] are therefore the more common approaches. From a Bayesian perspective, Tancredi and Liseo introduced a paradigm for probabilistic record linkage [104], and Steorts and coauthors proposed a Bayesian approach to graphical record linkage [102].

In the frequentist view, Fellegi and Sunter assumed that each record pair belongs to one of the two latent classes [47]. The distribution of comparison vector $\mathbf{\Gamma}$ for each pair is assumed to follow a mixture model, for $\boldsymbol{\gamma} \in \mathcal{L}$,

$$\begin{aligned}\mathbb{P}(\mathbf{\Gamma}_{i,j} = \boldsymbol{\gamma}_{i,j}) =& \mathbb{P}(\mathbf{\Gamma}_{i,j} = \boldsymbol{\gamma}_{i,j} \mid (a_i, b_j) \in M)\mathbb{P}((a_i, b_j) \in M) && (3.28)\\ &+ \mathbb{P}(\mathbf{\Gamma}_{\boldsymbol{i,j}} = \boldsymbol{\gamma}_{\boldsymbol{i,j}} \mid (a_i, b_j) \in U)\mathbb{P}((a_i, b_j) \in U). && (3.29)\end{aligned}$$

If we do not make additional assumptions on the joint agreement pattern, the comparison vector $\mathbf{\Gamma}$ may take 2^K different values, each of which corresponds to a parameter that we need to estimate. To reduce this number, some authors [e.g. 47, 116] have proposed to make the so-called conditional independence assumption between fields of the comparison vector. Under this assumption, we obtain :

$$\mathbb{P}\left[\mathbf{\Gamma}_{i,j} = (\gamma^1_{i,j}, \ldots, \gamma^K_{i,j}) | (a_i, b_j) \in M\right] = \prod_{k=1}^{K} \mathbb{P}(\Gamma^k_{i,j} = \gamma^k_{i,j} | (a_i, b_j) \in M), \quad (3.30)$$

$$\mathbb{P}\left[\mathbf{\Gamma}_{i,j} = (\gamma^1_{i,j}, \ldots, \gamma^K_{i,j}) | (a_i, b_j) \in U\right] = \prod_{k=1}^{K} \mathbb{P}(\Gamma^k_{i,j} = \gamma^k_{i,j} | (a_i, b_j) \in U). \quad (3.31)$$

The conditional independence assumption is common in most probabilistic record linkage models [116], although it may not hold in some practical cases. For example, if some records agree on a chronic disease, they are more likely to agree on the drug used. Although the assumption is invalid in some cases, the linkage result is still quite robust, in the sense that we may have a good linkage performance even if the conditional independence assumption does not hold [116, 51, 100]. Some authors [e.g. 118] relaxed this assumption

and showed better record linkage results in some specific scenarios.

Under the conditional independence assumption, we only need to estimate $2K + 1$ parameters which are the marginal probabilities of agreement for matched and unmatched pairs $m^k := \mathbb{P}(\gamma^k_{i,j} = 1|(a_i, b_j) \in M)$ and $u^k := \mathbb{P}(\gamma^k_{i,j} = 1|(a_i, b_j) \in U)$, and the overall matching probability $p_M := \mathbb{P}((a_i, b_j) \in M)$. Winkler proposed to apply the expectation maximization (EM) algorithm [116, 37, 117], to find the maximum likelihood estimates for the vector of parameters $\boldsymbol{\theta} := \left\{p_M, m^k, u^k, k = 1, \ldots, K\right\}$. It has become widely used in probabilistic record linkage [51, 20]. Once all the parameters are estimated, the record pairs may be ordered by either matching weights

$$\widehat{w}_{i,j} = \frac{\mathbb{P}(\boldsymbol{\Gamma}_{i,j} = \boldsymbol{\gamma}_{i,j}|(a_i, b_j) \in M, \widehat{\theta})}{\mathbb{P}(\boldsymbol{\Gamma}_{i,j} = \boldsymbol{\gamma}_{i,j}|(a_i, b_j) \in U, \widehat{\theta})},$$

see [47, 7], or by posterior probabilities of matching $\widehat{q}_{i,j} := \mathbb{P}((a_i, b_j) \in M|\boldsymbol{\Gamma}_{i,j}, \widehat{\theta})$ [76]. Then, the pairs are classified into matches, non-matches, or possible matches based on two defined thresholds [47]. Because the possible matches require manual review which is sometimes not available, Grannis and coauthors propose to establish only a single threshold to avoid human review [51]. Although the matching scores and the posterior probabilities produce the same ordering for record pairs [76], the posterior probabilities are preferable in our case because they may be useful for further analyses [73, 71, 62, 122].

In some applications, a one-to-one matching restriction may be needed; namely, each record in B can be matched to one and only one record in A, and conversely. One possible approach to respect a one-to-one matching is to solve a linear sum assignment problem proposed by [66]. If the optimal score is not demanded, a simple approach is to sort all candidate pairs according to their estimated posterior probabilities of matching and to select matched pairs in a greedy approach [20].

3.3.3 An extension of the Fellegi-Sunter model

In this section, we extend the Fellegi-Sunter model by making better use of low prevalence categorical matching variables and of continuous variables. Two new comparison approaches and a mixture model for mixed types of comparison values are introduced.

Comparison approaches

For a categorical matching variable, the proportions for each category are likely different, and accounting for these differences in a record linkage model may help to improve the linkage results. This idea was proposed by Fellegi and Sunter and Winkler [47, 115], and is applied to a real clinical data in [124]. These authors use the same model for simple agreement/disagreement comparison, but the matching weights are rescaled a posteriori, using a frequency-based correction. We introduce a new comparison approach for categorical matching variables, which differs from simple binary comparison and may naturally handle different proportions for categories.

Let X^k be a categorical matching variable taking L different values, which means that the comparison function for this variable may take up to L^2 values. For example, the comparison for a binary matching variable may lead to four possible realizations :

$$\{(0,0),(0,1),(1,0),(1,1)\}$$

and a comparison function can be defined as follows

$$h^k(0,0) = c_1, \quad h^k(0,1) = c_2, \quad h^k(1,0) = c_3 \quad \text{and} \quad h^k(1,1) = c_4, \tag{3.32}$$

where c_1, c_2, c_3 and c_4 stand for four different categories. It should be noted that the values taken by the comparison function have no ordinal meaning. If this is a low prevalence binary matching variable (e.g. a rare disease) such that only 5% (say) of the values in the dataset are equal to 1, the agreement on the value "1" is much more informative than the agreement on the value "0". Our comparison approach aims at using this information while the simple agreement comparison method does not, leading to poor performance. Hejblum and coauthors propose a Bayesian record linkage framework making use of a similar idea, which is efficient in the case of a large number of low-prevalence binary matching variables. However, their model is designed for binary variables only [58].

If the number of matching variables and/or the number of categories is large, the number of parameters to be estimated is $L^2 - 1$, which may be too large in practice. This number may be reduced by assigning the same comparison value for the agreement/disagreement of categories that have a close meaning. For instance, we may reduce the comparison

values given in (3.32) as

$$h^k(0,0) = c_1, \quad h^k(0,1) = h^k(1,0) = c_2 \quad \text{and} \quad h^k(1,1) = c_3. \tag{3.33}$$

In general, the number of comparison values depends on which realizations we would like to distinguish. Suppose that we are interested in a categorical matching variable X^k with categories $1, 2, \ldots, L$. If the first category seems particularly meaningful, we may distinguish whether we have an agreement on the first category, an agreement on another category, or a disagreement. In such case, the comparison function would be defined as

$$h^k(i,j) \quad = \quad \begin{cases} c_1 & \text{if } i = j = 1, \\ c_2 & \text{if } i = j \neq 1, \\ c_3 & \text{if } i \neq j = 1, \ldots, L. \end{cases}$$

The objective of this comparison approach is to distinguish the agreement of low prevalence values from other agreements, which differs from multiple levels of agreement introduced in [99] and [46].

Now, let us consider the case of a continuous variable X^k. For example, date variables (e.g., admission to the hospital, or medical act) are common in medical datasets. By converting each date into a duration from a specified origin, they may be treated as continuous counting variables. Even if an individual is present in both datasets, a lag between dates is likely to appear. The simple binary comparison is therefore not appropriate. In this work, if the k^{th} matching variable is continuous, we propose to consider

$$\Gamma^k_{i,j} = h^k(X^k_{A,i}, X^k_{B,j}) = d(X^k_{A,i}, X^k_{B,j}), \tag{3.34}$$

where d is a distance which can be used to measure the difference between two dates of events, in which case it can be interpreted as a time lag. By using the distance, the continuous comparison values $\gamma^k_{i,j}$ of matching pairs $(X^k_{A,i}, X^k_{B,j})$ can be described as

$$\Gamma^k_{i,j} | (a_i, b_j) \in M = \begin{cases} 0 & \text{with probability} \quad 1 - e^k, \\ \epsilon^k_{i,j} > 0 & \text{with probability} \quad e^k, \end{cases}$$

where e^k is the proportion of error, and $\epsilon^k_{i,j}$ is the error term of the k^{th} matching variable among matched pairs. For example, two patients who refer to the same individual should

have the same day for a medical act, up to some errors in the registration process, and the distance should therefore be equal to 0 or to a small error term $\epsilon_{i,j}^k$. Therefore, $\Gamma_{i,j}^k|(a_i, b_j) \in M$ follows a hurdle distribution in which the positive part depends only on the distribution of errors. On the other hand, the distribution of $\Gamma_{i,j}^k|(a_i, b_j) \in U$ depends mostly on the distribution of the k^{th} matching variable, since $\epsilon_{i,j}^k$ is often small compared to the distance between records for two unmatched units.

Estimation of parameters

Let

$$\boldsymbol{\Gamma}_{i,j} \quad = \quad \left(\Gamma_{i,j}^1, \ldots, \Gamma_{i,j}^{K_1}, \Gamma_{i,j}^{K_1+1}, \ldots, \Gamma_{i,j}^{K_1+K_2}\right) \tag{3.35}$$

be a mixed type comparison vector which includes K_1 categorical comparison values $\Gamma_{i,j}^1, \ldots, \Gamma_{i,j}^{K_1}$ and K_2 continuous distances $\Gamma_{i,j}^{K_1+1}, \ldots, \Gamma_{i,j}^{K_1+K_2}$. Following the Fellegi-Sunter framework, these comparison vectors are assumed to follow the mixture model (3.28).

Under the conditional independence assumption between the different fields in the comparison vector for both the matched and the unmatched sets, we have

$$\mathrm{f}_{\boldsymbol{\Gamma}_{i,j}|(i,j)\in M}(\boldsymbol{\gamma}_{i,j}) = \underbrace{\prod_{k=1}^{K_1} \mathbb{P}(\Gamma_{i,j}^k = \gamma_{i,j}^k|(a_i, b_j) \in M)}_{P_{i,j}^{1M}} \underbrace{\prod_{k=K_1+1}^{K_1+K_2} \mathrm{f}_{\Gamma_{i,j}^k|(a_i,b_j)\in M}(\gamma_{i,j}^k)}_{P_{i,j}^{2M}}, \tag{3.36}$$

$$\mathrm{f}_{\boldsymbol{\Gamma}_{i,j}|(i,j)\in U}(\boldsymbol{\gamma}_{i,j}) = \underbrace{\prod_{k=1}^{K_1} \mathbb{P}(\Gamma_{i,j}^k = \gamma_{i,j}^k|(a_i, b_j) \in U)}_{P_{i,j}^{1U}} \underbrace{\prod_{k=K_1+1}^{K_1+K_2} \mathrm{f}_{\Gamma_{i,j}^k|(a_i,b_j)\in U}(\gamma_{i,j}^k)}_{P_{i,j}^{2U}}, \tag{3.37}$$

for $i = 1, \ldots, n_A$ and $j = 1, \ldots, n_B$. For both equations (3.36) and (3.37), the first term in the right hand side involves K_1 categorical comparison values of the comparison vector $\boldsymbol{\Gamma}_{i,j}$. We define

$$m_s^k = \mathbb{P}(\Gamma_{i,j}^k = s|(a_i, b_j) \in M) \text{ and } u_s^k = \mathbb{P}(\Gamma_{i,j}^k = s|(a_i, b_j) \in U) \text{ for } s \in S^k, \tag{3.38}$$

with S^k the set of all possible categorical comparison values for the k^{th} variable. Then

$$P_{i,j}^{1M} = \prod_{k=1}^{K_1} \mathbb{P}(\Gamma_{i,j}^k = \gamma_{i,j}^k | (a_i, b_j) \in M) = \prod_{k=1}^{K_1} \prod_{s \in S^k} (m_s^k)^{\mathbb{1}\{\gamma_{i,j}^k = s\}},$$
$$P_{i,j}^{1U} = \prod_{k=1}^{K_1} \mathbb{P}(\Gamma_{i,j}^k = \gamma_{i,j}^k | (a_i, b_j) \in U) = \prod_{k=1}^{K_1} \prod_{s \in S^k} (u_s^k)^{\mathbb{1}\{\gamma_{i,j}^k = s\}},$$

for $i = 1, \ldots, n_A$ and $j = 1, \ldots, n_B$, and with $\sum_{s \in S^k} m_s^k = \sum_{s \in S^k} u_s^k = 1$.

The second part in the right hand side of equations (3.36) and (3.37) involves K_2 continuous values of the comparison vector $\mathbf{\Gamma}$. We define

$$\begin{aligned} P_{i,j}^{2M} &= \prod_{k=K_1+1}^{K_1+K_2} \mathrm{f}_{\Gamma_{i,j}^k | (a_i,b_j) \in M}(\gamma_{i,j}^k) \quad \text{with} \quad \mathrm{f}_{\Gamma_{i,j}^k | (a_i,b_j) \in M}(\gamma_{i,j}^k) = f_M^k(\phi_M^k), \\ P_{i,j}^{2U} &= \prod_{k=K_1+1}^{K_1+K_2} \mathrm{f}_{\Gamma_{i,j}^k | (a_i,b_j) \in U}(\gamma_{i,j}^k) \quad \text{with} \quad \mathrm{f}_{\Gamma_{i,j}^k | (a_i,b_j) \in U}(\gamma_{i,j}^k) = f_U^k(\phi_U^k), \end{aligned} \tag{3.39}$$

for $i = 1, \ldots, n_A$ and $j = 1, \ldots, n_B$. The distributions f_M^k and f_U^k need to be postulated, depending on the characteristics of the matching variables and on the chosen distance.

To find the maximum likelihood estimates for parameters, we apply the Expectation Maximization (EM) algorithm [37] or the Expectation Conditional Maximization (ECM) algorithm [84], depending on the distribution f^k.
Once all parameters are estimated using the EM/ECM algorithm, the posterior probabilities $q_{i,j} = \mathbb{P}((a_i, b_j) \in M | \mathbf{\Gamma}_{i,j} = \boldsymbol{\gamma}_{i,j})$ are estimated for all record pairs by the Bayes formula

$$\widehat{q}_{i,j} = \frac{\widehat{p}_M \widehat{P}_{i,j}^{1M} \widehat{P}_{i,j}^{2M}}{\widehat{p}_M \widehat{P}_{i,j}^{1M} \widehat{P}_{i,j}^{2M} + (1 - \widehat{p}_M) \widehat{P}_{i,j}^{1U} \widehat{P}_{i,j}^{2U}}. \tag{3.40}$$

These estimated posterior probabilities are then used to find proper matched pairs.

3.3.4 Conclusion

Proposed method. In this work, we have proposed an extension of the Fellegi and Sunter model, particularly when the matching data contains various types of variables. We have introduced a discrete distribution mixture model to handle categorical matching variables

with low prevalence and a hurdle gamma distribution mixture model to handle continuous matching variables.

Simulation results. The simulation studies show that our proposed model outperforms the simple model with binary comparison in all the scenarios considered. For categorical matching variables, we have shown that the proposed model is more efficient than the standard model, especially when there are low prevalence values. However, if the frequencies of the different categories of a matching variable are similar, then there is not much difference between our approach and the standard one. In that case, the model with binary comparison should be considered due to its simplicity. For continuous matching variables, the proposed mixture of hurdle gamma distributions performs better than the standard model and is robust to some misspecification of the distribution of the comparison function. However, our evaluation remains specific to the fact that we are dealing with continuous time variables, which may be naturally modeled by Gamma distributions. We also conducted a simulation with mixed-type data. Consistently with the previous simulation results, the proposed model has better performances than the standard model. In the application of real data, the performances are also better. We obtained a larger number of patients matched between the SNDS and the GETBO datasets, with high matching probabilities.

3.4 Perspectives

3.4.1 Potential developments for the OT algorithms

Heterogeneous domain adaptation with mixed type covariates. The use of Co-Optimal Transport is interesting when the covariate vectors in the two databases are of different dimensions. However, these algorithms require that all components of the covariate vector be defined on the same probability space, which is not the case when dealing with mixed-type covariates. A project would be to adapt this algorithm to mixed-type covariates.

Data recoding with different databases containing information collected at different times for the same individuals. In our previous work, the questionnaires to record the same information were different across different databases with different individuals. However, this problem can arise across different databases containing information collected at different times for the same individuals. So our model requires specific adaptations in this context.

Statistical modeling on imputed data after data merging. When performing a statistical analysis on the imputed data from the data merging process, the data distribution is modified. An objective would be to consider the data merging step to correct the estimations of the methods when performing a statistical analysis using imputed data. A predictive function could also be introduced in the models to combine the steps of data merging and statistical analysis.

Record linkage using optimal transportation. Optimal transport may be a tool for record linkage but algorithms still need to be developed.

3.4.2 Application to omics data

My work on data merging is part of a project in collaboration with IRSET whose research focuses on identifying the impact of the exposome in childhood on health outcomes.

Problem. The exposome represents the exposures to which a person is subjected throu-

ghout his/her life, including the chemical, microbiological, physical, recreational, and medicinal environments, lifestyle, diet, and infections. Pregnancy (prenatal period), childhood, and puberty have been identified as particularly sensitive periods for which environmental exposures may impact individual health trajectories. Lifecourse epidemiology needs to integrate data from multiple sources and increasingly complex markers of exposure and health effects that advances in biology, biochemistry, and bioinformatics have made available. Compiling large and multimodal data from different sources would allow to include more diverse populations and variable levels of exposure. The heterogeneity of the datasets is a major difficulty.

We are here interested in the detection of environmental stressors that could impact children's health outcomes. In this context, the mother-child cohorts TIMOUN (Guadeloupe, West French Indies, inclusions in 2004-2006), PELAGIE (Brittany, inclusions in 2002-2006), and EDEN (Poitiers, Nancy districts, France, inclusions in 2003-2006) are particularly interesting as they gather information on the effect of pre and postnatal environmental exposure to pollutants on children development and adolescents health.

However, the information on exposure can be collected using multiple matrices, multiple techniques, or multiple questionnaires between databases. Differences could be found between two cohorts that are pooled for a joint analysis, or within the same cohort, at different times of follow-up. So, there is a need to harmonize the exposure measure.

Multi-matrices exposure. Omics techniques (such as high-resolution mass spectrometry) are often applied to identify or even quantify the presence of pollutants in biological matrices such as urine, blood, and hair. The same chemical compounds could be detected in these different matrices. However, it often happens that, for some individuals, we don't have access to all the matrices. As an illustration, prenatal exposure to persistent organochlorine pollutants can be estimated in maternal blood during pregnancy or in umbilical cord blood at birth. When pooling several cohorts, one or both of the measurements may be available for all or a sub-sample of these cohorts.

Multi-techniques exposure. Untargeted chemical analyses are currently being developed and may exhibit the presence of different identified or unidentified compounds within the same biological matrix. Indeed, the laboratories have different techniques to analyze

the samples and, by consequence, some molecules are "quantified" differently, even in the same matrices. Furthermore, molecules may be not annotated so we can't identify which are common in the two files.

Multi-questionnaires exposure (or confounders). The questionnaires to record the same information can be different across different cohorts or in the same cohort at different times. The variables have different modalities and different numbers of modalities. As an example, children's behavior may be reported using different psychometric tools in two cohorts that are pooled or at different follow-ups of the same cohort. It can also be the case for some socio-demographic confounders such as the educational level of the mother, or the mode of daycare of the children that can be coded with different modalities.

Objectives. Our aim is to harmonize the exposure measures. Our OT algorithms and their possible adaptations introduced previously could deal with these different problems.

Chapitre 4

Survival analysis

My thesis and post-doctoral works are my first research projects in survival data analysis, where the proposed methodology was motivated by a problem encountered in a clinical trial (sections 1.2.1 and 1.2.2). The last part concerns the survival data analysis of matched data (section 4.3 introduced in section 2.1.3).

4.1 Survival analysis of preventive clinical trials

Problem. No effective curative treatment currently exists for Alzheimer disease, making its prevention a priority. During my PhD work, the rare published articles in the field of prevention trials for dementia which measured Alzheimer disease incidence as their primary outcome, had been negative [111]. The statistical analysis of these trials relies on the logrank test. This test is known to be optimal under the proportional hazards model, thus it may be inadequate for prevention clinical trials which may require a certain period of exposure to an intervention before an effect can be detected. The proportional hazards condition of optimality is unrealistic in this setting.

Objective. It is therefore conceivable to find more powerful tests that can capture non-proportional effects in order to improve survival analysis of the primary outcome in randomized controlled trials for the prevention of Alzheimer disease.

Proposed method. We proposed a methodology for using the Fleming-Harrington test, which allows for the detection of late effects. This is a family of tests depending on a parameter q. The setting of clinical trials imposes two constraints : the parameter q value must be fixed a priori and the sample size has to be calculated based on data similar to what we expect to observe. So, the alternative hypothesis for which the test is optimal for a fixed parameter q is studied, that gives a method for generating optimal data for Fleming-Harrington. The main result is that this test is not sensitive to a variation of this

parameter q which implies that variation on this specific parameter doesn't significantly influence the result of the test. Based on this, we proposed some tools for choosing an appropriate value of q in a given clinical trial and we provided a sample size formula for the Fleming-Harrington test for testing its optimal assumptions [VG6].

The constant piecewise test, which depends on a more easily explainable parameter, t, is better understood by methodologists. Indeed, this parameter can be seen as the moment when the late effects appear. A relationship between the parameters q and t was provided so that the Fleming-Harrington test with parameter q will be closest to the constant piecewise test with parameter t. This comparison shows that the best test is Fleming-Harrington [VG3].

Finally, the assumption of late treatment effects was relaxed. A new statistic defined by the maximum between the logrank statistic and Fleming-Harrington's statistic for a fixed parameter q was proposed. Obviously, this statistic is powerful enough under late and proportional effects. A method for calculating the required sample size for this test was provided. This performance of this statistic was compared to the Supremum statistic on time of a weighted logrank statistic [VG4].

4.2 Effect of a correlated competing risk on marginal survival estimation in an accelerated failure time model

Motivations :

- Competitive risks can arise when subjects may experience competing events, which censor the outcome of interest.
- Cox's partial likelihood estimator treating competing events as independent censoring is commonly used to examine group differences in clinical trials but fails to adjust for omitted covariates and can bias the assessment of marginal benefit.

Contributions of the Chapter [VG5] :

- Bivariate normal linear (BNV) model generating latent data with dependent censoring is used to assess this bias :
  - ✓ Assess the effects of correlation between two competing risks on the estimator of treatment effect in two-sample time-to-event studies in simulations with BVN competing risks.
  - ✓ Assess the robustness of the Cox model estimator of cause-specific hazard ratio to correlation when data are generated under this BNV model.
- A novel Expectation Maximisation algorithm.

Collaborations :

- Post doctorate work with *Malcolm Hudson* (Macquarie University, Sydney), *Val Gebski* (NHMR, University of Sydney) and *Maurizio Manuguerra* (Macquarie University, Sydney).

Package R : `bnc`.

4.2.1 Introduction

Problem and objective

In survival analysis problems, competitive risks can arise when a patient experiences an event that prevents the occurrence of the event of interest being observed. The role of covariates in competing risks is often studied by modeling the cause-specific hazard function [29] or the subdistribution hazard function [48]. In medical research, the Cox proportional

hazards model is one of the most commonly used approaches to study the relationship between variables and a time-to-event [29]. A competing risk occurring independently of the event of interest provides independent censoring. The partial likelihood estimator in Cox regression, modelling the individual cause-specific hazard function (which treats competing events as independent censoring) is commonly used to examine group differences in clinical trials. In this case, a coefficient estimated in Cox regression provides a hazard ratio (HR) for the marginal survival of interest (with or without treatment). However, a correlated competing risk introduces dependent censoring on the event of interest. When the correlation of competing risks is not effectively controlled by model variables, neglected covariates can induce unobservable correlation [70], toward which the Cox regression model is not robust [63]. The effect is that the estimated cause-specific HR may differ from the marginal HR [68, 44].

Parametric Accelerated Failure Time (AFT) models, such as the Weibull model, provide a valuable alternative to the proportional hazards (PH) assumptions [113]. In AFT models, bias in estimating marginal effects from regression coefficients is again anticipated when a correlated competing event censors follow-up.

This work is motivated by a study in which 143 patients have received surgery for metastatic melanoma. The trial compares two highly correlated events, the local relapse (event of interest) and the regional relapse (competing event) in a group of patients receiving adjuvant radiotherapy and in a control group [90].

Objective. The objective of this work is to evaluate the bias of the marginal hazard ratio estimator when competing events are correlated.

Literature

Bias in the assessment of marginal benefit with the presence of correlated competing risks. The estimated cause-specific HR may differ from the marginal HR [68, 44]. Emura *et al.* analyze this difference under copula-based dependent censoring [43]. They show that if the censoring probability is high, the difference is significant. Furthermore, the difference inflates as the dependence (copula) parameter deviates from zero. Lu *et al.* confirm in simulations that, in the presence of correlation, the estimated cause-specific

HRs for an event of interest can differ substantially from the marginal HR of the event of interest [80].

Why AFT models ? Hougaard notes that parametric AFT models are robust under certain conditions where PH models are not [63]. Klein *et al.* and Lambert *et al.* argue for parametric accelerated failure time models involving frailty-like terms as an effective alternative to parametric PH models [75, 72]. Weibull AFT models share the PH properties assumed by the semi-parametric Cox regression model, facilitating trial power and sample-size calculations under scenarios with varying treatment effects. In the alternative lognormal AFT model, the regression coefficient for the treatment indicator represents the treatment effect and its exponential provides the median ratio (MR, the ratio of median survival with and without treatment). Hence AFT models with lognormal marginals offer readily interpretable regression models for cause-specific median times to failure.

Our bivariate normal linear (BVN) AFT model is a specific case of the Deresa and Van Keilegom model. Recently, Deresa and Van Keilegom review parametric and semi-parametric approaches to regression modeling for competing risks and introduce a multivariate normal regression model for dependent censoring [39]. Their model allows for different forms of censoring (including loss to follow-up or termination of the study) and for an initial parameterized power transformation of survival time. Assumptions for model identifiability, including those required by the parameterized class of power transformations, are provided. The multivariate normal linear model of Deresa and Van Keilegom [38] is a more general formulation of our BVN censored linear model. This earlier model was applied in the simpler context of bivariate outcomes, identity power transformation, and covariates common to both causes.

Why parametric modelling ? In many applications involving competing risk data, individual events within a cluster may be correlated due to unobserved shared factors across individuals. Many authors are then concerned with non-parametric statistical tests for correlated competing risks data (see for example [15, 53]). Following Cox [interviewed in 95], we argue that non-parametric and semi-parametric modeling in competing risks can lead to overlooking the biases implicit in proportional hazards and independent censoring assumptions, while parametric models can add biological insight.

4.2.2 Linear survival models with dependent censoring

In this section, we establish notation and introduce the multivariate normal regression model of Deresa and Van Keilegom (the "DVK model") in the context of bivariate outcomes.

It is convenient in simulating an AFT model to measure times on a log scale and to restrict attention to the occurrence of an event of interest (event 1) or the occurrence of a competing event (event 2). While more than two events might be considered, interest in many medical trials is focused on a single primary outcome, with any competing events distracting from the measurement of treatment effect, with few observations to individually model some causes. Thus pooling any competing risks together is common. We restrict discussion to two event causes.

Since a competing risk is an event whose occurrence precludes the occurrence of the primary event of interest, only the first-occurring event is observed ; the observed time to an event is the minimum of two correlated times. Follow-up of events of types 1 and 2 are both subject to independent censoring at log-time C ; both event types are subject to the same censoring. This censoring is assumed to be non-informative. Therefore, we observe (Y_j^o, D_j) for $j = 1 \dots n$, with $Y_j^o = \min(Y_{j,1}, Y_{j,2}, C_j)$, and D_j indicating event type or censoring. Here

$$D_j = \begin{cases} 1: & \text{if the event of interest is observed,} \\ 2: & \text{if an event of a competing risk is observed,} \\ 0: & \text{if no event is observed during follow-up.} \end{cases}$$

The DVK model for competing risks

The study of the effect of correlation between competing events presupposes the existence of a joint distribution of latent variables. These are the latent failure times of each cause. However, this joint density is non-identifiable, so estimation of parameters from competing event data is not possible in a fully nonparametric context [see 107]. While identifiability is recovered by appropriate parametric assumptions, it remains important to evaluate the parametric model fit to observations.

First, consider the case of two competing events. Assume latent event times are pairs (Y_1, Y_2) with a bivariate Normal distribution, with a 2×2 covariance matrix $\mathbf{\Sigma}$ common to all subjects. Further assume each pair has expected values specified in a linear model from covariates $X_1, \ldots, X_p$. We refer to this as a BVN regression model for competing risks.

Specifically, assume that observations are an i.i.d. sample of size n from random vector $(Y_1, Y_2,\ X_1, \ldots, X_p,\ C)$, with the conditional distribution of (Y_1, Y_2), given $X_1 = x_1, \ldots, X_p = x_p$, being bivariate normal with mean vector $\boldsymbol{\mu} = (\boldsymbol{x}^\top \boldsymbol{\beta}_1, \boldsymbol{x}^\top \boldsymbol{\beta}_2)$ and covariance matrix $\mathbf{\Sigma}$, of dimension 2×2. Here Y_1 denotes the log-time to the event of interest $(k = 1)$ and Y_2 to an event of (any) other cause $(k = 2)$, with $\boldsymbol{x} = (x_1, \ldots, x_p)^\top$. Regression coefficients $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ form the columns of a $p \times 2$ matrix $\mathbf{B}$. Last, random variable C denotes the log-time to independent censoring. Assume hereafter that

1. $(Y_1, Y_2)^\top$ and C are conditionally independent given $(X_1, \ldots, X_p)^\top$, and that
2. $Y_1 - \mu_1, Y_2 - \mu_2,$ and C are independent of $(X_1, \ldots, X_p)^\top$.

The DVK model extends the bivariate normal model to the multivariate case and applies a parameterized transformation model to the log survival times of each cause. In it, covariates are specified to be cause-specific. Thus the DVK model is

$$\Lambda_\alpha(Y_k) = \boldsymbol{x}^\top \boldsymbol{\beta}_k + \epsilon_k, \tag{4.1}$$

for causes $k = 1, \ldots, m$, where $\Lambda_\alpha()$, is a generic parametric (α) class of monotone increasing transformations. The error vector (ϵ_1, ϵ_2) is bivariate normal, mean vector $\mathbf{0}$, covariance matrix $\mathbf{\Sigma}$.

In order to simplify notation, we confine discussion in the following sections to the BVN model $(m = 2)$ with all covariates for each cause in common and simple log transformation of survival times ($\Lambda : y \to y$ above). We use the more restrictive assumptions of the bivariate normal model to simplify notation in later sections; this is particularly appropriate when developing the Expectation Maximisation (EM) algorithm.

Cause-specific hazards of competing events

With data on the log scale, the cause-specific hazard function to a first-occurring event of cause 1 induced by the BVN joint distribution of Y_1, Y_2 is

$$\begin{aligned}\lambda_1(y; \mathbf{B}, \mathbf{\Sigma}) &= \lim_{dy \to 0} \mathbb{P}\left[Y_1 \in (y, y+dy) | Y_1 > y, Y_2 > y\right] / dy \\ &= \lim_{dy \to 0} \mathbb{P}\left[Y_1 \in (y, y+dy), Y_2 > y\right] / [\mathbb{P}\left[Y_1 > y, Y_2 > y\right] dy].\end{aligned}$$

Let $\Phi(z)$ denote the survival function of the univariate standard normal distribution and $\phi(z)$ the standard normal density. Let σ_1, σ_2 denote standard deviations of Y_1, Y_2. Define standardized values a, b of y under the two marginals by $a = (y - \mu_1)/\sigma_1$, $b = (y - \mu_2)/\sigma_2$, where $\mu_1 = \boldsymbol{x}^\top \boldsymbol{\beta}_1$ is the expectation of Y_1 and $\mu_2 = \boldsymbol{x}^\top \boldsymbol{\beta}_2$ is the expectation of Y_2. Then the numerator of the above expression for the hazard function becomes

$$\begin{aligned}\lim_{dy \to 0} \mathbb{P}\left[Y_1 \in (y, y+dy), Y_2 > y\right] / dy &= f_{Y_1}(y)\, \mathbb{P}\left[Y_2 > y \,|\, Y_1 = y\right] \\ &= \sigma_1^{-1}\, \phi(a)\, \Phi\left(\frac{b - \rho\, a}{\sqrt{1-\rho^2}}\right).\end{aligned}$$

Here $f_{Y_1}(y) = \sigma_1^{-1}\, \phi(a)$ is the univariate normal density of Y_1 and ρ is the correlation between Y_1 and Y_2. The final equality for the numerator follows because the conditional distribution of Y_2 given Y_1 is univariate normal :

$$[Y_2 | Y_1 = y] \;\sim \mathcal{N}\left(\mu_2 + \rho \frac{\sigma_2}{\sigma_1}(y - \mu_1),\; \sigma_2^2(1-\rho^2)\right).$$

Thus, defining the joint bivariate normal survival function as $S_{12}(y_1, y_2) = \mathbb{P}\left[Y_1 > y_1, Y_2 > y_2\right]$, we have

$$\lambda_1(y; \mathbf{B}, \mathbf{\Sigma}) = \sigma_1^{-1}\, \phi(a)\, \Phi\left(\frac{b - \rho\, a}{\sqrt{1-\rho^2}}\right) / S_{12}(y, y). \tag{4.2}$$

The corresponding hazard $\lambda_2(y; \mathbf{B}, \mathbf{\Sigma})$ for time to observing the competing risk is readily obtained by exchanging a and b. The sum of these hazards equals the hazard of time to the first event, their ratio determines the conditional probability of event type for first events at log time y.

Likelihood function and inference

In this section, we set notations for censoring outcomes. We then define a likelihood function for bivariate normal observations censored by a competing risk. In competing risks data all observations are subject to censoring, not only by the end of follow-up but also by competing events. When $D_j = 1$, time to event 2 is censored by an observed event of cause 1 ; i.e. $Y_{j,1} = y_j^o$ is observed and $Y_{j,2} > y_j^o$ is censored. Similarly, when $D_j = 2$, $Y_{j,2} = y_j^o$ is observed and $Y_{j,1} > y_j^o$. Finally, $D_j = 0$ when times to both events exceed the period of follow-up : $Y_{j,1} > y_j^o,\ Y_{j,2} > y_j^o$ for observed end time of follow-up $Y_j^o = y_j^o$. The distribution of $C_1, \ldots, C_n$ need not be included in likelihood calculations when censoring is independent and noninformative. Hence the likelihood function for competing risk observations $(Y_j^o = y_j^o, D_j = d_j;\ j = 1, \ldots, n)$ is defined as :

$$
\begin{aligned}
&L(\mathbf{B}, \mathbf{\Sigma}; \boldsymbol{y}^o, \boldsymbol{d}) \\
&\quad = \prod_{j:d_j=1} f_{Y_1}(y_j^o)\, \mathbb{P}\left[Y_2 > y_j^o \,|\, Y_1 = y_j^o\right] \times \prod_{j:d_j=2} f_{Y_2}(y_j^o)\, \mathbb{P}\left[Y_1 > y_j^o \,|\, Y_2 = y_j^o\right] \\
&\qquad \times \prod_{j:d_j=0} S_{12}(y_j^o, y_j^o), \\
&\quad = \prod_{j:d_j=1} f_{Y_1}(y_j^o)\, S_{2|1}(y_j^o|y_j^o) \prod_{j:d_j=2} f_{Y_2}(y_j^o)\, S_{1|2}(y_j^o|y_j^o) \prod_{j:d_j=0} S_{12}(y_j^o, y_j^o), \qquad (4.3)
\end{aligned}
$$

where $S_{1|2}(a|b) = \mathbb{P}[Y_1 > a \,|\, Y_2 = b]$, $S_{2|1}(a|b) = \mathbb{P}[Y_2 > a \,|\, Y_1 = b]$, and, as before, $S_{12}(a, b) = \mathbb{P}[Y_1 > a,\ Y_2 > b]$.

These probabilities depend on the design matrix $\boldsymbol{X}$, the coefficient matrix $\mathbf{B}$ and the covariance matrix $\mathbf{\Sigma}$ of the bivariate normal distribution. In the linear model, each observation has its own covariate values, a row vector of the covariate matrix $\boldsymbol{X}$. Let $\boldsymbol{M} = \boldsymbol{X}\mathbf{B}$ be the matrix of expected values of $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2)$ where $\mathbf{Y}_k = (Y_{1,k}, \ldots, Y_{n,k})^\top$ for $k = 1, 2$. Then $\boldsymbol{m}_k = \boldsymbol{X}\boldsymbol{\beta}_k$ is the vector containing expected times to event k. Define $\boldsymbol{z}_k = (\boldsymbol{y}^o - \boldsymbol{m}_k)/\sigma_k$, with $\sigma_k = \sqrt{\sigma_{k,k}}$, for cause $k = 1, 2$. Then $(\mathbf{Y} - \boldsymbol{M})\boldsymbol{W} \sim \mathrm{BVN}(\mathbf{0}, \boldsymbol{R})$ where the 2×2 weight matrix $\boldsymbol{W}$ is diagonal with entries $\sigma_1^{-1}, \sigma_2^{-1}$, and $\boldsymbol{R} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. Thus, given the covariate matrix $\boldsymbol{X}$, all likelihood terms may be expressed in terms of the standard bivariate normal distribution ; the likelihood is a function of ρ and values $\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{d}$ dependent on the observed random sample. $\boldsymbol{z}_1, \boldsymbol{z}_2$ are themselves functions of $\boldsymbol{y}^o$ and parameters $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1, \sigma_2$.

The well-known separability of the likelihood by competing risk [68, Ch.8] in AFT models

permits columns of $\mathbf{B}$ to be estimated individually after censoring events of other causes. However, this separability depends on parameters not being common to components of different causes. For the likelihood of the BVN model, some parameters, specifically those of $\mathbf{\Sigma}$, are involved in factors for each cause. Only in the case $\rho = 0$ will the optimization simplify. As noted above, a correlated BVN distribution assumption can be replaced by the lesser assumption of cause-specific hazards $\lambda_1(.), \lambda_2(.)$ defined in equation (4.2) in order to estimate $\mathbf{B}$ by Maximum Likelihood. However, in the correlated BVN distribution, these cause-specific hazards will be functions of $\mathbf{\Sigma}$, which we treat as unknown. Therefore, the optimization available for $\rho = 0$ is not readily available with correlated competing risks (even when the correlation is known).

Expectation Maximization algorithm

An EM algorithm may be employed to fit the BVN linear model parameters $\mathbf{B}, \mathbf{\Sigma}$. EM algorithms provide a steady assured convergence to ML or MPL solutions. In each iteration, an EM algorithm imputes sufficient statistics of missing data (latent times of each unobserved event), substituting expected values of sufficient statistics conditional on observed data and current parameter estimates (E step). New parameter estimates are then obtained from these imputed sufficient statistics as though they originated from a complete sample (the M-step, [37]). An EM algorithm for the likelihood function of survival Y_1, *not* subject to censoring by a competing risk, is described by Aitkin *op.cit.* For the likelihood function (4.3) for bivariate observations $(\boldsymbol{Y}^o, \boldsymbol{D})$, we introduce a new EM algorithm summarised below as Algorithm 3 :
Here, current estimates of the model parameters in iteration i are $\boldsymbol{\theta}^{(i)} = \left(\mathbf{B}^{(i)}, \mathbf{\Sigma}^{(i)}\right)$. The new estimate $\mathbf{B}^{(i+1)}$ of $\mathbf{B}$ is obtained as $\widehat{\mathbf{B}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \mathbf{Y}$ with imputed $\mathbf{Y}$. For this estimation, when $D_j = 2$, so that $Y_{j,1}$ is censored by the event of cause 2 at y_j^o, the censored observation $Y_{j,1}$ is imputed by $\mathbb{E}[Y_1 | Y_1 > y_j^o, Y_2 = y_j^o]$, where $\boldsymbol{M}^{(i)} = \boldsymbol{X}\mathbf{B}^{(i)}$. When $D_j = 0$, $Y_{j,1}$ is imputed by $\mathbb{E}[Y_1 | Y_1 > y_j^o, Y_2 > y_j^o; \boldsymbol{\theta}^{(i)}]$, with a similar expression for the imputation of $Y_{j,2}$. The expected values are calculated in each case by using the current iteration's matrix of means $\boldsymbol{M}^{(i)}$ and covariance matrix $\mathbf{\Sigma}^{(i)}$, and the conditional expectations. A complete data sufficient statistic for the covariance matrix $\mathbf{\Sigma}$ is $\mathbf{V} = (\mathbf{Y} - \boldsymbol{X}\widehat{\mathbf{B}})^\top (\mathbf{Y} - X\widehat{\mathbf{B}}) = \mathbf{Y}^\top \boldsymbol{Q}\mathbf{Y}$ for known projection matrix $\boldsymbol{Q} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$. The EM update to $\mathbf{\Sigma}$ therefore includes imputation of quadratic terms (squares and cross-products) in (Y_1, Y_2). The new estimate is obtained using the censored observation y_j^o by replacing linear terms as above and quadratic terms using appropriate conditional dis-

Algorithm 3: EM algorithm for BVN model

Initialisation : $\boldsymbol{\theta}^0 = (\boldsymbol{B}^0, \boldsymbol{\Sigma}^0)$. We used initial values : $\rho^{(0)} = 0$; $\beta_k^{(0)} = 0$, $\sigma_k^{(0)} = 1$, for $k = 1, 2$. Let $\boldsymbol{Q} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$.

for *i = 0,1,2, ...* **do**

  E-step : replace the unobserved variables with their conditional expectations given $\boldsymbol{Y}^o, \boldsymbol{D}, \boldsymbol{X}, \boldsymbol{C}$ at $\boldsymbol{\theta}^{(i)}$:

  if $D_j = 2$, $Y_{j,1} \leftarrow \mathbb{E}[Y_{j,1} \,|\, Y_{j,1} > y_j^o,\, Y_2 = y_j^o;\, \boldsymbol{\theta}^{(i)}]$, $Y_{j,1}^2 \leftarrow \mathbb{E}[Y_{j,1}^2 \,|\, Y_{j,1} > y_j^o, Y_{j,2} = y_j^o;\, \boldsymbol{\theta}^{(i)}]$,

  if $D_j = 1$, $Y_{j,2} \leftarrow \mathbb{E}[Y_2 \,|\, Y_1 = y_j^o,\, Y_2 > y_j^o;\, \boldsymbol{\theta}^{(i)}]$, $Y_{j,2}^2 \leftarrow \mathbb{E}[Y_{j,2}^2 \,|\, Y_{j,1} = y_j^o,\, Y_{j,2} > y_j^o;\, \boldsymbol{\theta}^{(i)}]$,

  if $D_j = 0$, $Y_{j,1} \leftarrow \mathbb{E}[Y_1 \,|\, Y_1 > y_j^o, Y_2 > y_j^o;\, \boldsymbol{\theta}^{(i)}]$, $Y_{j,1}^2 \leftarrow \mathbb{E}[Y_1 \,|\, Y_1 > y_j^o, Y_2 > y_j^o;\, \boldsymbol{\theta}^{(i)}]$, $Y_{j,2}^2 \leftarrow \mathbb{E}[Y_2 \,|\, Y_1 > y_j^o, Y_2 > y_j^o;\, \boldsymbol{\theta}^{(i)}]$, $Y_{j,1}Y_{j,2} \leftarrow \mathbb{E}[Y_{j,1}Y_{j,2} \,|\, Y_{j,1} > y_j^o, Y_{j,2} > y_j^o;\, \boldsymbol{\theta}^{(i)}]$

  M-step : Maximize the likelihood of sufficient statistics $\mathbf{X}^\top\mathbf{Y}$ and $\mathbf{Y}^\top\mathbf{Y}$ of equations (B2), (B3) of Appendix B in [VG5]. Calculate $\boldsymbol{B}^{(i+1)}, \boldsymbol{\Sigma}^{(i+1)}$ using equations (B1), (B4), (B5) to define $\boldsymbol{Y}^{(i)}$, then equations (B6) and (B7) :

  $\boldsymbol{Y}^{(i)} \leftarrow \mathbb{E}[\boldsymbol{Y} \,|\, \boldsymbol{Y}^o, \boldsymbol{D}, \boldsymbol{X}, \boldsymbol{C};\, \boldsymbol{\theta}^{(i)}]$

  $\boldsymbol{B}^{(i+1)} \leftarrow (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\mathbf{X}^\top\mathbf{Y}^{(i)}$

  $n\boldsymbol{\Sigma}^{(i+1)} \leftarrow \mathbb{E}\left[\mathbf{Y}^\top\mathbf{Y} \,|\, \boldsymbol{Y}^o, \boldsymbol{D}, \boldsymbol{X}, \boldsymbol{C};\, \boldsymbol{\theta}^{(i)}\right] - \boldsymbol{Y}^{(i)\top}\boldsymbol{Y}^{(i)} + \boldsymbol{Y}^{(i)\top}\boldsymbol{Q}\,\boldsymbol{Y}^{(i)}$

tributions. For $j' \neq j$ the statistical independence of observations reduces calculations to imputation of linear statistics, but for $j' = j$ more complex conditional expectations must be evaluated. For example, when $D_j = 1$, the quadratic term $Y_{j,1}^2$ is a known quantity. But, when $D_j = 2$, a quadratic term $Y_{j,1}^2$ is imputed using $\mathbb{E}[Y_{j,1}^2 | Y_{j,1} > y_j^o, Y_{j,2} = y_j^o]$. Similarly, when $D_j = 0$, the same term is imputed as $\mathbb{E}[Y_{j,1}^2 | Y_{j,1} > y_j^o, Y_{j,2} > y_j^o]$. When $D_j = 0$, we further require $\mathbb{E}[Y_{j,1}Y_{j,2} | Y_{j,1} > y_j^o, Y_{j,2} > y_j^o]$. Again these conditional expectations are calculated assuming current parameter values $\mathbf{B}^{(i)}, \boldsymbol{\Sigma}^{(i)}$, which provide the mean vector $(m_{j,1}^{(i)}, m_{j,2}^{(i)})^\top$. We provide all required moment results for the EM algorithm in a bivariate normal censored linear model in appendix A in [VG5], and include the corresponding R-code in the package `bnc`. Both the EM algorithm and these results for the bivariate normal distribution appear to be new and may prove useful in other contexts.

Standard errors of the EM algorithm solution are available using the numerical differentiation of Fisher scores (NDS) method of [65] and implemented in the R package `turboEM` [10]. The score statistic is derived from evaluations of the score function $Q(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$ which is accessible for computation using our probability results for the BVN complete data.

Mildly Penalize Likelihood

Convergence to a boundary point of the parameter space can introduce issues for optimization methods. Optimizers such as Newton's method break down near boundaries when iterations take them outside a constrained domain (such as $|\rho| \leq 1$). Specialized methods are available for constrained domains, but may affect the speed of convergence of the algorithm. In particular, the EM algorithm's rate of convergence is reduced when $\boldsymbol{\theta}$ lies on or near a boundary [88]. This is problematic when conducting large numbers of simulations. The rate of convergence is particularly affected in regions of flat likelihood. It is common to observe improvement after introducing a penalty term in the likelihood [52, 22]. Therefore, we regularize by penalizing log-Likelihood $\log L = \log L(\mathbf{B}, \boldsymbol{\Sigma}\,; y^o, \boldsymbol{d})$ as :

$$\log L \leftarrow \log L - (\nu + 3) \times \log(\det(\boldsymbol{\Sigma})/2 - \kappa \operatorname{tr}(\boldsymbol{\Sigma}^{-1})/2$$

where ν is a positive integer, the degrees of freedom parameter of the inverse-Wishart prior, $\boldsymbol{\Sigma} \sim \mathcal{IW}(\nu, \kappa\mathcal{I})$, whose posterior distribution provides the penalty term [122]. Higher degrees of freedom increasingly encourage a solution consistent with the prior mean, for which we specify a multiple (default $\kappa = 1$) of the identity matrix. We then utilize the `turboEM` algorithm for this MPL, using an accelerated EM form (method `squareEM`) [10].

4.2.3 Conclusion

Proposed method. In this study, we used a joint modeling approach for competitive events, enabling the consideration of potential correlations between these events. We have proposed a bivariate normal linear AFT model that generates latent data with dependent censoring to assess this bias. For the estimation of model parameters, the BVN model lends itself to the implementation of a novel Expectation Maximisation (EM) algorithm generalizing a similar algorithm for univariate survival proposed by Aitkin [3] This EM algorithm provides a solution for maximum likelihood estimation or penalized maximum likelihood estimation.

Simulation results. The parametric specification as a bivariate normal distribution together with likelihood penalization better regularizes an otherwise ill-posed problem, as demonstrated in simulations. This stabilization of parameter estimation, together with the model's ability to estimate marginal hazards and directly introduce correlation, are strengths of the strong parametric assumption.

Our simulations show the estimates of regression coefficients are reliable. These regression coefficients provide means and mean differences (relative differences when survival times are log-transformed). While the correlation estimation is unstable (even in sample size $n = 1000$) it exhibits little bias ; it may find use in validating external information on ρ.

Alternative Cox models are based on proportional hazards assumptions, which differ from those of AFT models. Moreover, with a positive correlation between competing risks our results on estimating marginal treatment-benefit show that Cox models are not robust to departures from a correlation $\rho = 0$. Thus the BVN model is a useful development for application when a strong correlation of the survival outcome with a competing risk is suspected, perhaps as a consequence of unmeasured covariates.

The BVN linear model assumes that the time-to-event distribution is lognormal. While the parametric form of hazards is difficult to identify from data, simulations with non-normal data suggest the model is robust. Our simulation findings concerning the lack of bias in regression parameter estimation are in accord with those of a different approach, which induces a correlation between latent event times through a copula model [17]. However, when the degree of association is misspecified in the copula, regression parameter estimates are severely biased. This indicates an advantage of the use of the BVN model, which provided reliable estimates by adapting to the level of association in the data.

Guidelines. The parametric model can be readily used as a sensitivity analysis for assessing the effects of correlation induced by neglected covariates. This method can introduce external information on the association of competing risks to assess its influence on other parameter estimates. Because of the uncertain ability to estimate ρ, our simulations considered an alternative to MPL estimates of treatment effect in the BVN model, fixing ρ to provide restricted estimates of other BVN model parameters.

Regression estimates of treatment effect may differ from corresponding Cox coefficients due to different models (accelerated failure time versus proportional hazard). When dependent censoring is present, and fitted parametric cumulative incidence functions agree with semi-parametric estimates, the BVN model can prove useful in estimating treatment effect for comparison with findings of alternative PH model fits.

4.3 Survival analysis of matched data

Motivations :

- Acounting for matching errors in survival data analysis when the record linkage process is not known.

$\Longrightarrow$ we don't have access to the matching variables and the individual matching probability estimates but we access an audit sample.

Contributions of the Chapter [VG13] :

- A bias-corrected estimating equation for Cox regression analysis with linked data and a variance estimator for the adjusted estimation of Cox regression coefficients.

Collaborations :

- PhD of *Huan Tanh Vo*. Co-direction with *Guillaume Chauvet* (ENSAI, Rennes), *Andre Happe* (Université de Rennes 1) and *Stephane Paquelet* (IRT bcom).

Perspectives :

- IAcounting for matching errors in survival data analysis when the record linkage process is known.
  PhD of *Vanessa Chezeu*. Co-direction with *Jean-François Dupuy* (INSA, Rennes) and *Samuel Bowong* (University of Douala, Cameroon).
- Package R.

4.3.1 Introduction

Problem and objective

Problem. The second part of *Huan Vo Tanh*'s PhD work concerns the statistical analysis of matched data when the variable of interest is a lifetime. Indeed, perfect matching is never achieved, and neglecting associated errors can lead to biased estimates. In most applications, the ultimate goal of record linkage is to obtain a linked dataset for a statistical analysis. Neter and his coauthors emphasized the potential biases associated with matching errors, which come in two forms : false links and missed links [87]. False links refer to records that are considered linked when they do not actually correspond to the same entity. On the other hand, missed links are records that refer to the same entity but are not identified as linked during the matching process. Since unique identifiers are not available or matching variables are likely to contain errors, false links are inevitable,

regardless of the record linkage method used. Therefore, the secondary analysts should be aware of linkage errors, and choose a suitable strategy to avoid biased estimators. In the absence of linkage errors, the estimation of coefficients in the Cox regression model is straightforward and approximately unbiased [4]. However, simulations demonstrate that even a small rate of linkage errors can lead to biased parameter estimation in the Cox model.

Objective. The objective is to account for linkage errors in the estimation of the parameters of the Cox proportional hazards model.

Specific problem. We consider the situation where explanatory variables and individuals' lifetimes are not reported in the same database (i.e. GETBO registry and SNDS data). We aim to conduct survival analyses in the dataset where we observe the lifetime variables (i.e. GETBO registry). However, we do not observe the true explanatory variables in this database but rather the variable obtained after linking with the SNDS dataset (see figure 4.1). In some cases, the persons performing the record linkage and the analysis may be the same, which is referred to as primary analysis. In other cases, the record linkage is performed by a trusted third party, and the person performing the statistical analysis has limited knowledge about the record linkage, and in particular the matching variables.

SNDS

		Matching variables			
	NIR	Postal code	Echodoppler date	Cancer	Exposure
a_1	~~1 96 09 99~~	29001	17/05/2013	0	No
a_2	~~2 93 08 88~~	29002	17/05/2013	0	No
a_3	~~2 85 06 66~~	29003	19/11/2013	0	Yes
a_4	~~1 99 05 55~~	29004	01/03/2014	0	No
⋮	⋮	⋮	⋮	⋮	⋮
a_m	~~1 86 03 65~~	29010	12/01/2015	0	No

GETBO registry

		Matching variables			Survival variables	
	NIR	Postal code	Echodoppler date	Cancer	Duration	AVC
b_1	~~1 96 09 99~~	29001	12/03/2014	1	7	0
b_2	~~2 93 08 88~~	29002	17/05/2013	0	9	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
b_n	~~1 90 01 86~~	29010	6/09/2016	1	15	0

GETBO registry

		Matching variables			Survival variables		
	NIR	Postal code	Echodoppler date	Cancer	Duration	AVC	Exposure
b_1	~~1 96 09 99~~	29001	12/03/2014	1	7	0	**Predicted**
b_2	~~2 93 08 88~~	29002	17/05/2013	0	9	1	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	
b_n	~~1 90 01 86~~	29010	6/09/2016	1	15	0	

FIGURE 4.1 – Analysis of linked survival data

Literature

Numerous authors [e.g. 73, 13, 122] have addressed this issue and proposed different methods to account for these linkage errors. However, these methods are mainly designed for linear and logistic regression models.

4.3.2 Cox regression analysis with linked data without knowledge of record linkage process

Cox regression model

The Cox proportional hazard model [29] is the most popular method to assess the effect of explanatory variables $\mathbf{X}$ on a survival time. This is therefore one of the most important models in medical research. Suppose that a random sample of n units is available. For each unit $i = 1, \ldots, n$, we let $\widetilde{T}_i$ be a non-negative random variable, which denotes the duration between a time origin and the time of occurrence of some event of interest. We suppose that $\widetilde{T}_i$ is right censored, which means that the event is observed only if it occurs before censoring time C_i. For units $i = 1, \ldots, n$, we therefore observe $T_i = \min(\widetilde{T}_i, C_i)$. We let $\delta_i = \mathbb{1}_{\{\widetilde{T}_i \leq C_i\}}$ denote the variable indicating whether the duration time is observed before censoring. We assume the censoring time C is non-informative of the survival parameter and independent of the true event time $\tilde{T}$. The vector of explanatory variables is denoted as $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,p})^T$. In this section, we first suppose that $\mathbf{X}_i$ is observed for any unit in the sample.

According to the Cox model, the hazard function of an event at time t is given by

$$\lambda(t|\mathbf{X}_i) = \lambda_0(t) \exp\left(\mathbf{X}_i^T \boldsymbol{\beta}\right), \tag{4.4}$$

where λ_0 is an unknown non-negative function of time (the so-called baseline hazard function) and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^\top \in \mathbb{R}^p$ is the unknown coefficient in the Cox PH. Let $\boldsymbol{\theta} = \{\Lambda_0, \boldsymbol{\beta}\}$, with $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$, is the set of parameters to be estimated. An estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is obtained by maximizing the partial likelihood, given by :

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{\exp(\boldsymbol{\beta}^\top \boldsymbol{X}_i)}{\sum_{k=1}^{n} Y_j(T_i) \exp(\boldsymbol{\beta}^\top \boldsymbol{X}_k)} \right)^{\delta_i}, \tag{4.5}$$

where $Y_j(t) = \mathbb{1}_{(T_j \geq t)}$ is an at-risk indicator [see for example 4]. Differentiating $\log L(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ yields the following estimating equation :

$$H_0(\boldsymbol{\beta}) := \frac{1}{n} \sum_{i=1}^{n} \delta_i \left\{ \mathbf{X}_i - \frac{\sum_{j=1}^{n} Y_j(T_i) \exp\left(\mathbf{X}_j^T \boldsymbol{\beta}\right) \mathbf{X}_j}{\sum_{j=1}^{n} Y_j(T_i) \exp\left(\mathbf{X}_j^T \boldsymbol{\beta}\right)} \right\} = 0, \tag{4.6}$$

We call (4.6) the theoretical estimating equation. The solution of this equation is called the maximum partial likelihood estimator of $\boldsymbol{\beta}$. It is consistent and asymptotically normal [4]. Under some mild assumptions, a consistent estimator of the covariance matrix of $\widehat{\boldsymbol{\beta}}$ is given by

$$\widehat{\mathbb{V}}_{\text{mpl}}(\widehat{\boldsymbol{\beta}}) = \left\{ -n \nabla H_0(\widehat{\boldsymbol{\beta}}) \right\}^{-1}, \tag{4.7}$$

see [4].

Linkage error model

Suppose that we have a dataset A of n_A time-to-event data. If the explanatory variables $\mathbf{X}_i$ were known for any unit $i \in A$, the parameter of the Cox model would be estimated by solving the theoretical estimating equation (4.6). However, if the explanatory variables are not known in database A, equation (4.6) may not be solved in practice.

In order to obtain the needed explanatory variables, a linkage is performed with a dataset B of size $n_B \geq n_A$, containing in particular the auxiliary variables $\mathbf{X}_i$. For any unit i in A, we note $\mathbf{Z}_i$ for the vector of auxiliary values resulting from the linkage process. Reasoning from the secondary analysis perspective, we do not have access to the matching variables and do not know the actual linkage process.

We assume that the linkage error is non-informative of the regression model, i.e. may depend on the errors in the matching process, but not on the model explanatory variables nor on the survival time [e.g. 14]. This is the key assumption of most secondary analysis approaches in the literature, for which Zhang and Tuoto have proposed a diagnostic

test [121]. Adopting the modeling approach in [23], we suppose that both databases are partitioned into blocks A_v and B_v, $v = 1, \ldots, V$, and that the record linkage is performed independently in these blocks. Also, we suppose that for any entity $i \in A_v$, we have :

$$\mathbf{Z}_i = \begin{cases} \mathbf{X}_i & \text{with probability } \alpha_v, \\ \mathbf{X}_{(j)} & \text{with probability } 1 - \alpha_v, \end{cases} \tag{4.8}$$

where (j) stands for some unit randomly selected in database B_v. In other words, it is supposed that for any $i \in A_v$, the correct entity is linked to i with probability α_v, otherwise, the unit j linked to i is randomly selected in B_v. We suppose that the linkage is performed independently for any unit $i \in A_v$, conditionally on the $\mathbf{X}_j$'s for $j \in B_v$.

It should be noted that we implicitly assume that A is a subset from B and that all entities in A can therefore have some matching records in B. Also, we assume that there is at most one link for each record of both databases. We assume that the linkage is complete, or alternatively that any missing links are independent of the time of event and model explanatory variables.

Adjusted estimating equation

By naively treating the linked explanatory variables $\mathbf{Z}_i$ as if they were the true explanatory variables $\mathbf{X}_i$ for the units $i \in A$, an estimator of $\boldsymbol{\beta}_0$ may be obtained by solving the following equation :

$$H_{\text{naive}}(\boldsymbol{\beta}) := \frac{1}{n_A} \sum_{v=1}^{V} \sum_{i \in A_v} \delta_i \left\{ \mathbf{Z}_i - \frac{\sum_{v=1}^{V} \sum_{j \in A_v} Y_j(T_i) \exp(\mathbf{Z}_j^\top \boldsymbol{\beta}) \mathbf{Z}_j}{\sum_{v=1}^{V} \sum_{j \in A_v} Y_j(T_i) \exp(\mathbf{Z}_j^\top \boldsymbol{\beta})} \right\} = 0. \tag{4.9}$$

We call (4.9) the naive estimating equation. Since some units are incorrectly linked, it may lead to biased estimates.

We propose a bias-corrected estimating equation, accounting for the fact that from the hit-miss model (4.8), the explanatory variables may be incorrectly linked. We first introduce some notations. Let us define

$$g(\boldsymbol{\beta}, \mathbf{X}_i) = \exp(\mathbf{X}_i^\top \boldsymbol{\beta}) \quad \text{and} \quad h(\boldsymbol{\beta}, \mathbf{X}_i) = \exp(\mathbf{X}_i^\top \boldsymbol{\beta}) \mathbf{X}_i.$$

Also, let $\bar{\mathbf{X}}_{B_v}, \bar{g}_{B_v}(\boldsymbol{\beta})$ and $\bar{h}_{B_v}(\boldsymbol{\beta})$ denote the means of $\mathbf{X}_i, g(\boldsymbol{\beta}, \mathbf{X}_i)$ and $h(\boldsymbol{\beta}, \mathbf{X}_i)$ over B_v, respectively. The linkage-error *adjusted estimating equation* (AEE) is given by

$$\bar{H}(\boldsymbol{\beta}) := \frac{1}{n_A} \sum_{v=1}^{V} \sum_{i \in A_v} \delta_i \left\{ \mathbf{X}_i^*(\alpha_v) - \frac{\sum_{v=1}^{V} \sum_{j \in A_v} Y_j(T_i) h_j^\star(\alpha_v, \boldsymbol{\beta})}{\sum_{v=1}^{V} \sum_{j \in A_v} Y_j(T_i) g_j^\star(\alpha_v, \boldsymbol{\beta})} \right\} = 0, \tag{4.10}$$

where, for any $i \in A_v$,

$$\begin{aligned} \mathbf{X}_i^\star(\alpha_v) &= \alpha_v^{-1} \mathbf{Z}_i - (\alpha_v^{-1} - 1) \bar{\mathbf{X}}_{B_v}, \\ g_j^\star(\alpha_v, \boldsymbol{\beta}) &= \alpha_v^{-1} g(\mathbf{Z}_j, \boldsymbol{\beta}) - (\alpha_v^{-1} - 1) \bar{g}_{B_v}(\boldsymbol{\beta}), \\ h_j^\star(\alpha_v, \boldsymbol{\beta}) &= \alpha_v^{-1} h(\mathbf{Z}_j, \boldsymbol{\beta}) - (\alpha_v^{-1} - 1) \bar{h}_{B_v}(\boldsymbol{\beta}). \end{aligned} \tag{4.11}$$

We prove that $\bar{H}(\boldsymbol{\beta})$ is an (approximately) conditionally unbiased estimator for the function $H_0(\boldsymbol{\beta})$ involved in the theoretical estimating equation.

Since there is no closed-form solution for the estimating equations considered above, an iterative method like the Newton-Raphson algorithm is commonly used in practice. Also, the probabilities α_v may be (somewhat arbitrarily) specified by the record linkage practitioner, or estimated from an audit sample [13, 121] if their true values are unknown.

Variance estimator

In this section, we discuss variance estimation for the estimator of the parameter β_0 obtained by solving the AEE given in (4.10). We first note that several sources of variance need to be accounted for a) the (usual) variability associated with solving a sample-based estimating equation, b) the variability associated with the linkage process, and c) the variability associated with the estimation of the probabilities $\alpha_v,\ v = 1, \ldots, V$. Using the variance estimator given in (4.7) fails to account for all these sources of variability, and therefore leads to an underestimation of the variance (see the simulation study in [VG13]).

We propose a sandwich-like variance estimator, which reads as follows :

$$\widehat{\mathbb{V}}_{\text{AEE}}(\widehat{\boldsymbol{\beta}}) := \{\nabla \bar{H}(\widehat{\boldsymbol{\beta}})\}^{-1} \times \widehat{\mathbb{V}}\{\bar{H}(\boldsymbol{\beta}_0)\} \times \{\nabla \bar{H}(\widehat{\boldsymbol{\beta}})\}^{-1}, \tag{4.12}$$

$$\text{with } \widehat{\mathbb{V}}\{\bar{H}(\boldsymbol{\beta}_0)\} = \widehat{\mathbb{V}}_1\{\bar{H}(\boldsymbol{\beta}_0)\} + \widehat{\mathbb{V}}_2\{\bar{H}(\boldsymbol{\beta}_0)\}. \tag{4.13}$$

The first component $\widehat{\mathbb{V}}_1\{\bar{H}(\boldsymbol{\beta}_0)\}$ in (4.13) accounts for the variability in (c). Under the

assumption that the validation samples S_v used for such estimation are selected in the datasets A_v through simple random sampling without replacement, this variance estimator is

$$\widehat{\mathbb{V}}_1\{\bar{H}(\boldsymbol{\beta}_0)\} = \sum_{v=1}^{V} \bar{H}_{2,v}(\widehat{\alpha}_v, \widehat{\boldsymbol{\beta}})\{\bar{H}_{2,v}(\widehat{\alpha}_v, \widehat{\boldsymbol{\beta}})\}^\top \times \left(\frac{1}{n_{S_v}} - \frac{1}{n_{A_v}}\right) \frac{n_{S_v}}{n_{S_v}-1} \frac{1-\widehat{\alpha}_v}{\widehat{\alpha}_v^3},$$

where n_{S_v} is the sample size of the validation set S_v, and

$$\bar{H}_{2,q}(\alpha_v, \boldsymbol{\beta}) = \frac{1}{n_A} \sum_{i \in A_v} \delta_i \{(\mathbf{Z}_i - \bar{\mathbf{X}}_{B_v}) - \frac{\sum_{j \in A_v} Y_j(T_i) \left\{\left\{h(\boldsymbol{\beta}, \mathbf{Z}_j) - \bar{h}_{B_v}(\boldsymbol{\beta})\right\} - R_i^*(\alpha_v, \boldsymbol{\beta}) \left\{g(\boldsymbol{\beta}, \mathbf{Z}_j) - \bar{g}_{B_v}(\boldsymbol{\beta})\right\}\right\}}{\sum_{j \in A_v} Y_j(T_i) g_j^\star(\alpha_v, \boldsymbol{\beta})}\}.$$

with

$$R_i^\star(\alpha_v, \boldsymbol{\beta}) = \frac{\sum_{j \in A_v} Y_j(T_i) h_j^\star(\alpha_v, \boldsymbol{\beta})}{\sum_{j \in A_v} Y_j(T_i) g_j^\star(\alpha_v, \boldsymbol{\beta})}.$$

The second component $\widehat{\mathbb{V}}_2\{\bar{H}(\boldsymbol{\beta}_0)\}$ in (4.13) accounts for both the variability in (a) and (b). We have

$$\widehat{\mathbb{V}}_2\{\bar{H}(\boldsymbol{\beta}_0)\} = \frac{s_H^2(\widehat{\boldsymbol{\beta}})}{n_A}$$

where

$$s_H^2(\boldsymbol{\beta}) = \frac{1}{n_A - 1} \sum_{v=1}^{V} \sum_{i \in A_v} \left\{ H_i(\boldsymbol{\beta}) - \frac{1}{n_A} \sum_{v=1}^{V} \sum_{j \in A_v} H_j(\boldsymbol{\beta}) \right\}^2$$

and

$$H_i(\boldsymbol{\beta}) = \delta_i \left\{ \mathbf{X}_i^\star(\widehat{\alpha}_v) - \frac{\sum_{v=1}^{V} \sum_{j \in A_v} Y_j(T_i) h_j^\star(\widehat{\alpha}_v, \boldsymbol{\beta})}{\sum_{v=1}^{V} \sum_{j \in A_v} Y_j(T_i) g_j^\star(\widehat{\alpha}_v, \boldsymbol{\beta})} \right\}.$$

4.3.3 Conclusion

Proposed methodology. By adopting the hit-miss linkage error model [23], we have proposed an adjusted estimating equation for Cox regression with matched data. This model aims to correct the bias present in the naive approach that ignores errors arising from

false links. We also introduce a variance estimator for the estimated regression coefficient. This variance estimator captures the entire variability, including that arising from linkage errors. In practice, accessing a random audit sample is necessary to estimate the errors introduced by false links. In this audit sample, we know whether the predicted links are correct or not, as the actual matching variables are not available.

Simulation results. Our simulations proved that the naive use of linked data may lead to substantial bias in a Cox regression model. Through various simulation scenarios with one block and also multiple blocks, the proposed adjusted estimating equation is shown to lead to substantial bias reductions as compared to the naive estimating equation.

4.4 Some methodological guidelines for epidemiology

4.4.1 The minimum number of subjects at risk required to interpret a Kaplan-Meier curve

Problem and objective. Still in the field of survival analysis, a frequently asked question by clinicians is the acceptable number of subjects at risk to interpret a Kaplan-Meier curve, which represents the estimation of the survival function of an event of interest.

Proposed methodology. In this work, we proposed a reliable index to determine the amount of information required for constructing Kaplan-Meier curves over time [VG8]. This index guides the user on the minimum number of subjects at risk needed at any given time in the study to ensure a reasonable interpretation of Kaplan-Meier estimation at different time points. It also allows the user to understand up to what time point a Kaplan-Meier curve can be extended. The construction of the index is based on both the magnitude of the decrease in Kaplan-Meier estimation as an additional event occurs over time and the variability of the estimation if all subjects had been followed until the point of interest.

Let us consider a sample of N subjects followed over time until the event of interest occurs. Let $\widehat{S}(t)$ be the Kaplan-Meier survival probability estimator at time t when $n(t)$ subjects remain at risk. At time $t = 0$, $n(0) = N$, $S(0) = 1$, and $\widehat{S}(t)$ decreases over time as events occur. If an additional event had occurred immediately after time t, the decrease in the estimated percentage of subjects not experiencing an event would be $\Delta(t)$, where

$$\Delta(t) = 100\widehat{S}(t)/n(t)$$

is defined for $n(t) \geq 1$. We refer to $\Delta(t)$ as the sensitivity index of the survival estimate at time t. If the estimated survival probability $\widehat{S}(t)$ at time t is high and only a few subjects remain at risk, the sensitivity index will be substantial, indicating the potential for a large gap in the Kaplan-Meier curve due to a single additional event.

We propose that the decrease in survival estimate at a time due to the occurrence of an additional event does not exceed the width of the 95% one-sided confidence interval for

survival based on "complete information," which is given by

$$\Delta(t) = 100 \times \widehat{S}(t)/n(t) \leq 100 \times 1.645 \times \sqrt{\widehat{S}(t)[1 - \widehat{S}(t)]/N}$$

which leads to

$$n(t) \geq \frac{1}{1.645}\sqrt{\frac{N\widehat{S}(t}{1 - \widehat{S}(t)}}.$$

We interpret this as evidence that enough subjects remain at risk at time t for the meaningful interpretation of the Kaplan-Meier plot.

The approach is applied to various epidemiological studies. The proposed index is straightforward to compute, provides guidance to users on how to extend the Kaplan-Meier curve (graphically), and prevents premature conclusions from being drawn due to insufficient information at any given time.

4.4.2 Study of the association between air pollution exposure and cancer risk

We provided guidance on the specific methodology of this study.

Problem. Black carbon (BC), a component of fine particulate matter (particles with an aerodynamic diameter of less than or equal to 2.5 μm ($PM_{2.5}$)), may contribute to the carcinogenic effects of air pollution. This study aimed to estimate the associations between long-term exposure to $PM_{2.5}$ and the risk of cancer.

Data. In this study [VGa7], the influence of different methods for assessing exposure to $PM_{2.5}$ on cancer risk estimates is examined in the large French Gazel cohort, a population-based cohort with 20,625 participants at enrollment, followed for 26 years, and with complete residential histories. Two endpoints of cancer incidence were investigated : all cancers combined and lung cancer. Two distinct methods for assessing $PM_{2.5}$ exposure were employed : a land use regression (LUR) model for Western Europe (covering the period from 1990 to 2015), and a chemistry-dispersion model (Gazel-Air) for France (covering the period from 1999 to 2015), each with a time series of annual concentrations spanning at least 20 years.

Proposed method.

Use of cumulative exposure to BC.

The exposure decreasing over time and cancer incidence increasing with time, we could observe a positive effect of exposure on cancer incidence. Therefore, we chose to consider cumulative exposure.

We utilized the extended time-dependent Cox model, with age as the time scale and cumulative exposure to BC as the time-dependent covariate, adjusted for sociodemographic and lifestyle variables. This model expresses the instantaneous risk function of developing cancer at time t as a function of time-dependent exposure $X(t)$ and sociodemographic and lifestyle variables $\boldsymbol{Z}$:

$$\lambda(t|X(t), \boldsymbol{Z}) = \lambda_0(t) \exp\left(\beta_1 X(t) + \boldsymbol{\beta}_2^\top \boldsymbol{Z}\right),$$

where β_1 and $\boldsymbol{\beta}_2$ represents the coefficients.

To account for the latency between exposure and cancer diagnosis, we applied a 10-year lag. Therefore, we considered only cases diagnosed between 1999 and 2015, with corresponding exposures occurring 10 years prior to incidence/censoring.

Use of piecewise Cox model due to nonlinearity.

To test the non linearity, we used spline functions with three degrees of freedom to represent the exposure in the Cox model (without adjustment) :

$$X(t) = \sum_{j=1}^{p} \alpha_j \varphi_j(t)$$

with $\varphi_1, \ldots, \varphi_p$ a B-spline base and α_j, $j = 1, \ldots, p$ coefficients. Figure 4.2 depicts the corresponding HRs of cancer and their corresponding confidence intervals concerning cumulative exposure, using the lowest exposure as a reference, for both exposure assessment methods. The "slope" HRs on the left side of the plot illustrate the slope of the curve

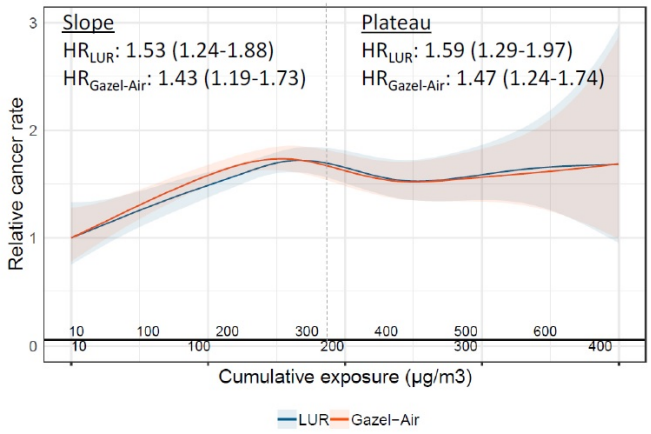

FIGURE 4.2 – Associations between cumulative $PM_{2.5}$ exposure estimated using either LUR or Gazel-Air and the risk of all types of cancer in the Gazel cohort.

below the 65th percentile of exposure (dashed vertical line) at 315 $\mu g/m^3$ for LUR and 185 $\mu g/m^3$ for Gazel-Air. These HRs are expressed for an increase of one interquartile range (IQR) in cumulative $PM_{2.5}$ exposure (IQR = 216 $\mu g/m^3$ for LUR, 127 $\mu g/m^3$ for Gazel-Air). On the right side, the "plateau" HRs represent the plateau above the 65th percentile of exposure, calculated for the 80th percentile of each exposure assessment method (380 $\mu g/m^3$ for LUR, 215 $\mu g/m^3$ for Gazel-Air).

Due to this nonlinearity, we modeled the relationship with a piecewise linear model, introducing an interaction term between $PM_{2.5}$ exposure and a binary variable indicating whether exposure was below or above the 65th percentile. The model is written as :

$$\lambda(t|X(t), \boldsymbol{Z}) = \lambda_0(t) \exp\left(\beta_1 \mathbb{1}_{\{X(t)<s\}} X(t) + \beta_2 \mathbb{1}_{\{X(t)\geq s\}} X(t) + + \boldsymbol{\beta}_3^\top \boldsymbol{Z}\right),$$

with s the threshold correspond to the 65th percentile of the exposure $X(t)$, β_1, β_2 and $\boldsymbol{\beta}_3$ represent the coefficient.

4.5 Perspectives

4.5.1 Acounting for matching errors in survival data analyses with a known linkage process

Specific problem. In the context of *Vanessa Chezeu*'s PhD, we consider a scenario where the linkage process is known, and we have access to the matching variables as well as the individual matching probability estimates. These probabilities allow us to link each patient in the SNDS to her/his most likely counterpart in the health registry. These probabilities also convey some uncertainty in the matching process, and this uncertainty must be taken into account in any subsequent statistical analysis.

Proposed methodology. Let's recall the notations. Suppose that we have a dataset A of n_A time-to-event data. Recall that we do not observe the explanatory variables $\boldsymbol{X}_i$. In order to obtain the needed explanatory variables, a linkage is performed with a dataset B of size $n_B \geq n_A$, containing in particular the auxiliary variables $\mathbf{X}_i$. We know that the explanatory variable vector for individual i takes one value from the set $\{\boldsymbol{x}_1, \cdots \boldsymbol{x}_{n_B}\}$ in B. For any individual i in A, we note $\mathbf{Z}_i$ for the vector of auxiliary values, that will be affected to the individual, resulting from the linkage process. We wish to estimate $\boldsymbol{\beta}$ in model (4.4) described in section 4.3.2. From the record linkage process described in chapter 3.3, we know that

$$\mathbb{P}(\boldsymbol{Z}_i = \boldsymbol{x}_j) = q_{i,j}, \quad j = 1, \ldots, n_B,$$

where

$$q_{i,j} = \mathbb{P}((a_i, b_j) \in M \mid \boldsymbol{\Gamma}_{i,j} = \boldsymbol{\gamma}_{i,j}), \quad i = 1, \ldots, n_A \text{ and } j = 1, \ldots, n_B \tag{4.14}$$

are the posterior probabilities of matching introduced in section 3.3.2. We define the normalized probabilities

$$p_{i,j} = \frac{q_{i,j}}{\sum_{j=1}^{n_B} q_{i,j}},$$

so that they sum to 1 for $j = 1, \ldots, n_B$. We proposed several estimation methods for the parameter $\boldsymbol{\beta}$.

A naive approach

According to the record linkage process, for each unit $i \in A$, the choice of the $\boldsymbol{Z}_i$ value depends on the posterior probabilities $\{q_{i,j}, j = 1, \ldots, n_B\}$ estimated for each comparison pair. We propose to replace $\boldsymbol{X}_i$ by the surrogate $\boldsymbol{Z}_i$, defined as :

$$\boldsymbol{Z}_i = \boldsymbol{x}_j \quad \text{where} \quad j = \ \text{argmax}_{1 \leq j \leq n_B}(q_{i,j}).$$

By naively treating the linked explanatory variables $\boldsymbol{Z}_i$ as the true explanatory variables $\boldsymbol{X}_i$, the partial likelihood in this case is obtained by replacing the values of $\boldsymbol{X}_i$ by the surrogate variable $\boldsymbol{Z}_i$ in the equation (4.5) from which we can estimate $\boldsymbol{\beta}$ by an estimator $\boldsymbol{\beta}_{\text{naive}}$ obtained by solving the estimating equation (4.6) by replacing $\boldsymbol{X}_i$ by $\boldsymbol{Z}_i$.

To reduce the bias of this naive estimator, we proposed several approaches to take into account the probabilistic aspect of the linked explanatory variables $\boldsymbol{Z}_i$.

Weighted average approaches

We first propose that $\boldsymbol{Z}_i$ is the average of all the available explanatory variables $\boldsymbol{x}_j$, $j = 1, \ldots, n_B$, weighted by the posteriors matching probabilities $p_{i,j}$ such as :

$$\boldsymbol{Z}_i = \sum_{j=1}^{n_B} p_{i,j} \boldsymbol{x}_j.$$

The partial likelihood in this case is obtained by replacing the values of $\boldsymbol{X}_i$ by the surrogate variable $\boldsymbol{Z}_i$ in the equation (4.5). The estimator $\boldsymbol{\beta}_{\text{W},1}$ of $\boldsymbol{\beta}$ is obtained by solving the following equation :

$$H_{\text{W},1}(\boldsymbol{\beta}) := \sum_{i=1}^{n_A} \delta_i \left\{ \sum_{j=1}^{n_B} p_{i,j} \boldsymbol{x}_j - \frac{\sum_{k=1}^{n_A} \sum_{j=1}^{n_B} Y_k(T_i) p_{k,j} \exp\left(\boldsymbol{\beta}^\top \boldsymbol{x}_j\right) \boldsymbol{x}_j}{\sum_{k=1}^{n_A} \sum_{j=1}^{n_B} Y_k(T_i) p_{k,j} \exp\left(\boldsymbol{\beta}^\top \boldsymbol{x}_j\right)} \right\} = 0.$$

For individual i, $i = 1, \ldots, n_A$, let $p_{i,j_1(i)}$ and $p_{i,j_2(i)}$ be the highest and the second highest elements of the vector $(p_{i1}, \ldots, p_{in_B})^\top$ respectively :

$$p_{i,j_1} = p_{i,j_1(i)} = \max_{1 \leq j \leq n_B} p_{i,j} \quad \text{and} \quad p_{i,j_2} = p_{i,j_2(i)} = \max_{j \neq j_1; 1 \leq j \leq n_B} p_{i,j}.$$

and $\boldsymbol{Z}_{j_1(i)}$ and $\boldsymbol{Z}_{j_2(i)}$ the j_1 and j_2 elements in the vector $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_B})^\top$. Let $p^\star_{i,j_k(i)} = p_{i,j_k(i)}/(p_{i,j_1(i)} + p_{i,j_2(i)})$ for $k = 1, 2$. We propose to replace $\boldsymbol{X}_i$ by $\boldsymbol{Z}_i = p^\star_{i,j_1(i)}\boldsymbol{Z}_{j_1(i)} + p^\star_{i,j_2(i)}\boldsymbol{Z}_{j_2(i)}$. Similarly to the previous method, an estimator $\boldsymbol{\beta}_{\mathrm{W},2}$ of $\boldsymbol{\beta}$ is obtained by solving the estimating equation :

$$H_{\mathrm{W},2}(\boldsymbol{\beta}) := \sum_{i=1}^{n_A} \delta_i \left\{ \left(p^\star_{i,j_1(i)}\boldsymbol{Z}_{j_1(i)} + p^\star_{i,j_2(i)}\boldsymbol{Z}_{j_2(i)} \right) - \frac{\sum_{k=1}^{n_A} Y_k(T_i) \left(p^\star_{k,j_1(k)} \exp\left(\boldsymbol{\beta}^\top \boldsymbol{Z}_{j_1(k)}\right) \boldsymbol{Z}_{j_1(k)} + p^\star_{k,j_2(k)} \exp\left(\boldsymbol{\beta}^\top \boldsymbol{Z}_{j_2(k)}\right) \boldsymbol{Z}_{j_2(k)} \right)}{\sum_{k=1}^{n_A} Y_k(T_i) \left(p^\star_{k,j_1(k)} \exp\left(\boldsymbol{\beta}^\top \boldsymbol{Z}_{j_1(k)}\right) + p^\star_{k,j_2(k)} \exp\left(\boldsymbol{\beta}^\top \boldsymbol{Z}_{j_2(k)}\right) \right)} \right\} = 0.$$

Complete likelihood with unobserved variables

Let $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_B}\}$ be the set of possible values of $\boldsymbol{Z}_i$ and $\boldsymbol{\theta} = \{\Lambda_0, \boldsymbol{\beta}^\top\}$. If $\boldsymbol{Z}_i$ were observed, we could write the following likelihood, based on the observations $\{T_i, \delta_i, \boldsymbol{Z}_i\}, i = 1, \ldots, n_A$:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n_A} \mathbf{f}_{(T,\delta,\boldsymbol{Z})}\left(T_i, \delta_i, \boldsymbol{Z}_i, \boldsymbol{\theta}\right).$$

As in the usual Cox model, this likelihood does not have a maximum in Λ_0 so we need to redefine a set of parameters. Let be the set of distinct event times $I = \{\tilde{t}_1, \tilde{t}_2, \ldots, \tilde{t}_D\}$ where, $D = \sum_{i=1}^{n_A} \delta_i$ with $D \leq n_A$. The continuous cumulative function $\Lambda_0(t)$ can be approximated by a step function with jumps at the observed uncensored event times [11]. Let us define the jump size of the step function at the different event times $\tilde{t}_d$ by $\Lambda_0\{\tilde{t}_d\}$. With these notations, the step function at time t is defined as :

$$\sum_{d=1}^{D} \Lambda_0\{\tilde{t}_d\} \mathbb{1}_{\left\{\tilde{t}_d \leq t\right\}}. \tag{4.15}$$

So, we redefine the set of parameters as $\boldsymbol{\theta} = \{\Lambda_0\{\tilde{t}_1\}, \ldots, \Lambda_0\{\tilde{t}_D\}, \boldsymbol{\beta}^\top\}$.

Here, $\boldsymbol{Z}_i$ is not observed. According to Fisher's identity, the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_{n_A}$ satisfy :

$$\mathbb{E}_{\boldsymbol{Z}|T,\delta}\left[\frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta}) \mid_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_n} | (T_i, \delta_i), 1 \leq i \leq n_A, \widehat{\boldsymbol{\theta}}_{n_A} \right] = 0. \tag{4.16}$$

We have :

$$\begin{aligned}\mathbb{E}_{\boldsymbol{Z}|T,\delta}\left(\log L(\boldsymbol{\theta}) \mid (T_i,\delta_i), 1\leq i\leq n_A, \widehat{\boldsymbol{\theta}}_{n_A}\right)&\\ &= \sum_{i=1}^{n_A}\mathbb{E}_{\boldsymbol{Z}|T,\delta}\left(\log\left(\mathbf{f}_{(T,\delta,\boldsymbol{Z})}\left(T_i,\delta_i,\boldsymbol{Z}_i,\boldsymbol{\theta}\right)\right) \mid T_i,\delta_i,\widehat{\boldsymbol{\theta}}_{n_A}\right),\\ &= \sum_{i=1}^{n_A}\sum_{j=1}^{n_B}\log\left(\mathbf{f}_{(T,\delta,\boldsymbol{Z})}\left(T_i,\delta_i,\boldsymbol{x}_j,\boldsymbol{\theta}\right)\right)\mathbb{P}\left(\boldsymbol{Z}_i=\boldsymbol{x}_j \mid T_i,\delta_i,\widehat{\boldsymbol{\theta}}_{n_A}\right), \qquad (4.17)\end{aligned}$$

where

$$\log\left(\mathbf{f}_{(T,\delta,\boldsymbol{Z})}\left(T_i,\delta_i,\boldsymbol{x}_j,\boldsymbol{\theta}\right)\right)=\log\left[\mathbf{f}_{(T,\delta|\boldsymbol{Z})}\left(T_i,\delta_i \mid \boldsymbol{x}_j,\boldsymbol{\theta}\right)\mathbb{P}\left(\boldsymbol{Z}_i=\boldsymbol{x}_j\right)\right]$$

and

$$\mathbb{P}\left(\boldsymbol{Z}_i=\boldsymbol{x}_j \mid T_i,\delta_i,\widehat{\boldsymbol{\theta}}_{n_A}\right)=\frac{\mathbf{f}_{(T,\delta|\boldsymbol{Z})}\left(T_i,\delta_i \mid \boldsymbol{x}_j,\widehat{\boldsymbol{\theta}}_{n_A}\right)\mathbb{P}\left(\boldsymbol{Z}_i=\boldsymbol{x}_j\right)}{\sum_{k=1}^{n_B}\mathbf{f}_{(T,\delta|\boldsymbol{Z})}\left(T_i,\delta_i \mid \boldsymbol{x}_k,\widehat{\boldsymbol{\theta}}_{n_A}\right)\mathbb{P}\left(\boldsymbol{Z}_i=\boldsymbol{x}_k\right)}.$$

Since C is non-informative of the survival parameter and independent of $\tilde{T}$, the likelihood of the data (T_i,δ_i) knowing the observation $\boldsymbol{x}_j$ is given by :

$$\begin{aligned}\mathbf{f}_{(T,\delta|\boldsymbol{Z})}\left(T_i,\delta_i \mid \boldsymbol{x}_j,\boldsymbol{\theta}\right) &= (f_{\tilde{T}}(T_i \mid \boldsymbol{x}_j,\boldsymbol{\theta})S_C(T_i))^{\delta_i}(S_{\tilde{T}}(T_i \mid \boldsymbol{x}_j,\boldsymbol{\theta})f_C(T_i))^{1-\delta_i},\\ &\propto f_{\tilde{T}}(T_i \mid \boldsymbol{x}_j,\boldsymbol{\theta})^{\delta_i}S_{\tilde{T}}(T_i \mid \boldsymbol{x}_j,\boldsymbol{\theta})^{1-\delta_i},\end{aligned}$$

where $S_{\tilde{T}}(T_i \mid \boldsymbol{x}_j,\boldsymbol{\theta}) = \exp\left(-\Lambda_0(T_i)\exp(\boldsymbol{\beta}^\top\boldsymbol{x}_j)\right)$ and $f_{\tilde{T}}(T_i \mid \boldsymbol{x}_j,\boldsymbol{\theta}) = \lambda_{\tilde{T}}(T_i \mid \boldsymbol{x}_j,\boldsymbol{\theta})S_{\tilde{T}}(T_i \mid \boldsymbol{x}_j,\boldsymbol{\theta})$.

The EM algorithm consists of maximizing (4.17) and alternates between two steps : the expectation and maximization steps. Let $(\boldsymbol{\theta}^{(r)})^\top = \{\Lambda_0^{(r)}\{\tilde{t}_1\},\ldots,\Lambda_0^{(r)}\{\tilde{t}_D\},(\boldsymbol{\beta}^{(r)})^\top\}$ is obtained in the r-th iteration of the EM algorithm.

Expectation step. The expectation step consists on calculating the value $\pi_{i,j}(\boldsymbol{\theta}^{(r)})$, equals to :

$$\begin{aligned}\pi_{i,j}(\boldsymbol{\theta}^{(r)}) &= \mathbb{P}\left(\boldsymbol{Z}_i=\boldsymbol{x}_j \mid T_i,\delta_i,\boldsymbol{\theta}^{(r)}\right)\\ &= \frac{p_{i,j}\left[\exp\left(\boldsymbol{\beta}^{(r)\top}\boldsymbol{x}_j\right)\right]^{\delta_i}\exp\left(-\Lambda_0^{(r)}(T_i)\exp(\boldsymbol{\beta}^{(r)\top}\boldsymbol{x}_j)\right)}{\sum_{k=1}^{n_B}p_{i,k}\left[\exp\left(\boldsymbol{\beta}^{(r)\top}\boldsymbol{x}_k\right)\right]^{\delta_i}\exp\left(-\Lambda_0^{(r)}(T_i)\exp(\boldsymbol{\beta}^{(r)\top}\boldsymbol{x}_k)\right)},\end{aligned}$$

with

$$\Lambda_0^{(r)}(T_i) = \sum_{d=1}^{|D|} \Lambda_0^{(r)}\{\tilde{t}_d\} \mathbb{1}_{\left\{\tilde{t}_d \leq T_i\right\}}.$$

Maximisation step. The maximisation step consists of calculating $\boldsymbol{\theta}^{(r+1)}$. The baseline hazard $\Lambda_0^{(r+1)}\{\tilde{t_d}\}$ is expressed as, $\forall d = 1, \ldots, \mid D \mid$,

$$\Lambda_0^{(r+1)}\{\tilde{t_d}\} = \frac{1}{\sum_{k=1}^{n_A} \sum_{\ell=1}^{n_B} \pi_{k,\ell}(\boldsymbol{\theta}^{(r)}) \exp(\boldsymbol{\beta^{(r)}}^\top \boldsymbol{x}_\ell) \mathbb{1}_{\left\{\tilde{t}_d \leq T_k\right\}}} \tag{4.18}$$

and $\boldsymbol{\beta}^{(r+1)}$ can be obtained by solving the following estimating equation :

$$H(\boldsymbol{\beta}) := \sum_{i=1}^{n_A} \delta_i \left(\sum_{j=1}^{n_B} \pi_{i,j}(\boldsymbol{\theta}^{(r)}) \boldsymbol{x}_j - \frac{\sum_{k=1}^{n_A} \sum_{\ell=1}^{n_B} 1_{\{T_i \leq T_k\}} \pi_{k,\ell}(\boldsymbol{\theta}^{(r)}) \exp\left(\boldsymbol{\beta}^\top \boldsymbol{x}_\ell\right) \boldsymbol{x}_\ell}{\sum_{k=1}^{n_A} \sum_{\ell=1}^{n_B} 1_{\{T_i \leq T_k\}} \pi_{k,\ell}(\boldsymbol{\theta}^{(r)}) \exp(\boldsymbol{\beta}^\top \boldsymbol{x}_\ell)} \right) = 0. \tag{4.19}$$

First simulations. This last approach, based on the maximization of the complete likelihood, seems to correct the bias present in the three other approaches. These developed techniques are more appropriate for applications to health data from the SNDS where the linkage process is known.

4.5.2 Other modelling

I will continue this work in collaboration with researchers at the institute IRSET. The follow-ups of the different cohorts introduced in section 3.4.2 become a huge challenge when children become older, and attrition may prevent from analysing long-term health effects (cardio-metabolic health, as an example) of early exposures that were collected at inclusion in the cohorts. The linkage of these cohorts with the SDNS may be a way to follow the health trajectories of the children without contacting them. The matching with the SNDS may also allow collecting health data that were not asked about in the cohort questionnaires and studying them as health outcomes or confounders in other association studies. We may also introduce matching errors in other survival models than the Cox model.

In our previous work, we considered the situation where the explanatory variables are reported in the SNDS while the individuals' lifetimes are reported in the cohort in which we wish to conduct survival analysis. However, in this application, the individuals' lifetimes

may be reported in the SNDS and the explanatory variables in the cohort (see table 4.1). So, methods need to be developed in this context. We may also propose a joint model in which the estimation of matching probabilities and Cox model parameters are performed simultaneously. We may also account for matching errors in survival models other than the Cox model. In the long term, we will develop a package to make our methods accessible.

SNDS

		Matching variables			Survival variables	
	NIR	Postal code	Echodoppler date	Cancer	Duration	AVC
a_1		29001	17/05/2013	0	7	0
a_2		29002	17/05/2013	0	9	1
a_3		29003	19/11/2013	0	10	1
a_4		29004	01/03/2014	0	12	0
⋮		⋮	⋮	⋮	⋮	⋮
a_m		29010	12/01/2015	0	6	1

Cohort

		Matching variables			
	NIR	Postal code	Echodoppler date	Cancer	Exposure
b_1		29001	12/03/2014	1	No
b_2		29002	17/05/2013	0	
⋮		⋮	⋮	⋮	No
b_n		29010	6/09/2016	1	No

TABLE 4.1 – Analysis of linked data. explanatory variables are observed in the SNDS and survival variables are observed in the database where we wish to perform the analyses.

Chapitre 5

Some contributions for medical data analysis

5.1 Causal analysis

Motivations :

- In observational cohort studies, confounding may occur when the distribution of baseline covariates differs between treated and control subjects.
- The propensity score is one of the methods that helps in reducing or minimizing this confounding to get valid inferences on treatment effects. The propensity score is generalized into a propensity function for quantitative exposures which is known as the generalized propensity score.
- Considering this type of exposure variable, one may be interested in estimating the dose-response function (treatment effect). In this case, no valid closed-form variance of the dose-response function estimator has been proposed.

Contributions of the Chapter [VG1] :

- Developing and evaluating closed-form variance estimators for stratified and weighted treatment effect estimators using the influence function linearization technique.

Collaborations :

- *David Hajage* (APHP, Paris) and *Guillaume Chauvet* (ENSAI, Rennes)

5.1.1 Introduction

Problem and objective

Problem. In observational cohort studies, confounding may occur when the distribution of baseline covariates differs between treated and control subjects (see figure 5.1). Propensity score (PS) methods help in reducing or minimizing this confounding and are widely used

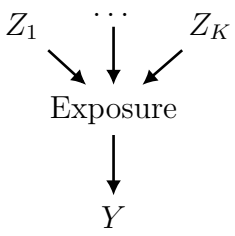

FIGURE 5.1 – Directed acyclic graph. $Z_1, \dots Z_K$ are the covariates and Y is the variable of interest.

in observational medical studies to assess the marginal effects of treatment or exposure on an outcome variable. The PS is the probability of receiving treatment and is first estimated, typically through logistic regression on the available covariates. The PS is then used to define weights for treated and untreated individuals (PS weighting methods) and allows for estimating both the average treatment effect on the overall population (ATE) and the average treatment effect on the treated population (ATT). In the case of two treatment groups, prior research [55] has shown that the variance of the ATE and ATT is accurately estimated only when the variance estimator considers that the propensity score is not known but rather estimated from the available data in a preliminary analysis stage. In many studies, the exposure of interest is continuous rather than binary. For example, we may not only know whether an individual is a smoker or not, but also the pack-years of cigarettes smoked, or the duration of smoking. Another example is the body mass index, which may be more informative as a continuous variable than if reduced to a dummy variable indicating obesity [123, 5]. The Generalized Propensity Score (GPS) is an extension of the propensity score, historically developed for binary exposures, to be used with quantitative or continuous exposures.

Objective. When dealing with this type of exposure variable, one might be interested in estimating the dose-response function (DRF) and estimating the variance of this estimator.

Literature

The propensity score has been generalized into a propensity function for quantitative exposures which is known as the generalized propensity score (GPS) [61, 64, 9, 50, 123, 5]. Similarly to the binary case, different propensity score methods have been proposed to estimate the treatment effects on outcomes using the GPS : covariate adjustment, [5] stratification [123] and inverse probability of treatment weighting (IPTW) [123, 5].

In the case of binary exposure, several authors have proposed valid closed-form variance estimators adapted to each treatment effect estimators : adjustment,[125] stratification, [114] matching, [1] and weighting [82, 81, 54]. Note that all these estimators take into account the fact that the theoretical propensity score value of an individual is unknown, and is estimated from the data in the first stage of analysis. However, variance estimation for treatment effects estimated using the GPS framework has received little attention.

5.1.2 Treatment effect estimator using the generalized propensity score

Notations and assumptions

Let T denote the level of a quantitative exposure which is a continuous variable, and $\boldsymbol{Z}$ a set of K baseline measured covariates. Let $Y(t),\ t \in \Psi$, denote a set of potential outcomes which is assumed to exist under Rubin's framework for causal inference. More precisely, we assume that T is a continuous exposure (i.e., Ψ is a subset of $\mathbb{R}$) and that $Y_i(t)$ is the outcome that would be observed for subject i if he/she received (maybe contrary to the reality) the level of exposure $T = t$. In practice, we only observe one level of exposure for each subject i and the corresponding outcome. The observed data consists of $(\boldsymbol{Z}_i, T_i, Y_i)$ for subjects $i = 1, \ldots, n$. We are interested in estimating the dose-response function

$$\mu(t) \quad = \quad \mathbb{E}[Y_i(t)], \tag{5.1}$$

which corresponds to the average response if all subjects were exposed to the level $T = t$. In randomized studies, it can be assumed that $Y(t)$ is independent of T, which is denoted as $Y(t) \perp\!\!\!\perp T, \quad \forall t \in \Psi$. In this work, we only assume that $Y(t)$ is independent of T given

$\boldsymbol{Z}$, which is known as the weak unconfoundedness assumption and denoted as

$$Y(t) \perp\!\!\!\perp T|\boldsymbol{Z}, \quad \forall t \in \Psi. \tag{5.2}$$

This assumption means that any association between the actual exposure and the potential outcomes is explained by a set of baseline covariates $\boldsymbol{Z}$ [61, 123]. Note that this assumption cannot be checked from the data.

Let us denote by

$$r(t \mid \boldsymbol{z}) \;:=\; f_{T|\boldsymbol{Z}}(t|\boldsymbol{z}), \tag{5.3}$$

the conditional density of exposure variable T given the covariates, which is called the generalized propensity score (GPS) [61]. We make the positivity assumption, namely

$$r(t \mid \boldsymbol{z}) > 0 \quad \text{for any} \quad t \in \Psi \text{ and for any } \boldsymbol{z}. \tag{5.4}$$

This means that any level of exposure $T = t$ is possible for any subject, whatever his/her baseline characteristics. A violation of this assumption may lead to biased estimators, or estimators with a large variability [86]. Note that this assumption may and should be checked from the data.

We focus on two approaches for the estimation of the dose-response function : inverse probability of treatment weighting, and stratification. Both approaches are presented in Zhang *et al.*,[123] and are briefly described in Sections 5.1.2 and 5.1.2.

Weighted treatment effect estimator

The first estimator is obtained by fitting a generalized linear regression model between the dose-response function and the exposure, used as the sole dependent variable. We considered a generalized linear model for the dose-response function, namely :

$$g\{\mu(T)\} \;=\; \beta_0 + \beta_1\, T, \tag{5.5}$$

where g is the link function. The estimator $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1)^\top$ is obtained as the solution of the estimating equation

$$U(\boldsymbol{\beta}) \quad = \quad \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \times \widehat{w}_i \{Y_i - \mu_i\}$$

with $\mu_i = \mu(T_i)$ and the weights $\widehat{w}_i$ that we use are presented thereafter. Focusing on the linear case (link function $g(x) = x$), our model is

$$\mu(T) \quad = \quad \beta_0 + \beta_1 \, T, \tag{5.6}$$

where β_0 is the average response observed in the case of null exposure ($T = 0$), and β_1 is the average response change if the level of exposure is increased by one unit. Other dose-response functions (e.g. in case of non-linear relationship) and/or other link functions may be better suited for other types of outcome (e.g., a binomial link function for a binary outcome), and may therefore be alternatively used.

The parameter $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$ is estimated by weighted least squares, which leads to

$$\widehat{\boldsymbol{\beta}}_w \quad := \quad (\widehat{\beta}_{w,0}, \widehat{\beta}_{w,1})^\top = \left(\sum_{i=1}^{n} \widehat{w}_i \tilde{T}_i \tilde{T}_i^\top \right)^{-1} \left(\sum_{i=1}^{n} \widehat{w}_i \tilde{T}_i Y_i \right) \tag{5.7}$$

where $\tilde{T}_i = (1, T_i)^\top$.This leads to the first estimator

$$\widehat{\mu}_w(t) \quad = \quad \widehat{\beta}_{w,0} + \widehat{\beta}_{w,1} \, t \tag{5.8}$$

for the dose-response function.

The weights used in equation (5.7) are the estimations of the theoretical Generalized Propensity Score (GPS) weights, defined as

$$w_i(\boldsymbol{\gamma}) \quad = \quad \frac{W(T_i|\boldsymbol{\gamma})}{r(T_i|\boldsymbol{Z}_i, \boldsymbol{\gamma})}, \tag{5.9}$$

where $r(t|\boldsymbol{z}, \boldsymbol{\gamma})$ is the conditional density of the exposure variable defined in (5.3), and where $W(\cdot|\boldsymbol{\gamma})$ is a stabilization factor. As is currently done in the literature, we use $W(t|\boldsymbol{\gamma}) := f_T(t|\boldsymbol{\gamma})$ the marginal density of the exposure variable. Note that the weights

depend on some unknown vector of parameters $\boldsymbol{\gamma}$, which needs to be estimated.

We suppose that T_i follows a normal distribution, both conditionally on $\boldsymbol{Z}_i$ and non conditionally. We may therefore write

$$f_T(t|\boldsymbol{\gamma}) = \frac{1}{\sqrt{2\pi\sigma_T^2}} \exp\left\{-\frac{1}{2\sigma_T^2}\left(T_i - \mu_T\right)^2\right\}, \tag{5.10}$$

$$r(t|\boldsymbol{z},\boldsymbol{\gamma}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left(T_i - \boldsymbol{\alpha}^\top \tilde{\boldsymbol{Z}}_i\right)^2\right\}, \tag{5.11}$$

with $\tilde{\boldsymbol{Z}}_i^\top = (1, \boldsymbol{Z}_i^\top)$ and $\boldsymbol{\gamma} = (\mu_T, \sigma_T^2, \boldsymbol{\alpha}^\top, \sigma^2)^\top$. The parameters μ_T and σ_T^2 in equation (5.10) are estimated by

$$\widehat{\mu}_T = \frac{1}{n}\sum_{i=1}^{n} T_i \quad \text{and} \quad \widehat{\sigma}_T^2 = \frac{1}{n-1}\sum_{i=1}^{n}(T_i - \widehat{\mu}_T)^2. \tag{5.12}$$

By fitting a linear regression model between the exposure variable and the covariates, namely

$$T_i = \tilde{\boldsymbol{Z}}_i^{\top}\boldsymbol{\alpha} + \eta_i, \tag{5.13}$$

the parameters $\boldsymbol{\alpha}$ and σ^2 in equation (5.11) are estimated by

$$\widehat{\boldsymbol{\alpha}} = \left(\sum_{i=1}^{n} \tilde{\boldsymbol{Z}}_i \tilde{\boldsymbol{Z}}_i^\top\right)^{-1}\left(\sum_{i=1}^{n} \tilde{\boldsymbol{Z}}_i T_i\right), \tag{5.14}$$

$$\widehat{\sigma}^2 = \frac{1}{n-K-1}\sum_{i=1}^{n}(T_i - \widehat{\alpha}^T \tilde{\boldsymbol{Z}}_i)^2.$$

This leads to the estimator $\widehat{\boldsymbol{\gamma}} = (\widehat{\mu}_T, \widehat{\sigma}_T^2, \widehat{\boldsymbol{\alpha}}^\top, \widehat{\sigma}^2)^\top$. By plugging this estimator in (5.9), we obtain the estimated weights $\widehat{w}_i := w_i(\widehat{\boldsymbol{\gamma}})$ used in equation (5.7). The model (5.13) is called the propensity model.

Stratified treatment effect estimator

The weighted estimator of the dose-response function considered in equation (5.8) of Section 5.1.2 proceeds through a linear regression on the whole sample, using weights to adjust for possible imbalance in the covariates.

An alternative approach consists in partitioning the sample into L strata, in such a way that the units inside a given stratum are somewhat similar with respect to the covariates. This may be done by fitting the propensity model in (5.13), ordering the units in the sample with respect to the prediction $\boldsymbol{Z}_i^\top \widehat{\alpha}$, and using the quantiles as cut-off points [123].

Inside any stratum $\ell = 1, \ldots, L$, we fit the regression model

$$Y(T) \quad = \quad \beta_{\ell,0} + \beta_{\ell,1}\, T + \epsilon_\ell, \tag{5.15}$$

and by estimating the parameter $\boldsymbol{\beta}_\ell = (\beta_{\ell,0}, \beta_{\ell,1})^\top$ by ordinary least squares, we obtain

$$\widehat{\boldsymbol{\beta}}_\ell \quad := \quad (\widehat{\beta}_{\ell,0}, \widehat{\beta}_{\ell,1})^\top = \left(\sum_{i \in S_\ell} \tilde{T}_i \tilde{T}_i^\top \right)^{-1} \left(\sum_{i \in S_\ell} \tilde{T}_i Y_i \right), \tag{5.16}$$

with S_ℓ the subset of sampled units that belong to the stratum ℓ. The stratified estimator of the parameter $\boldsymbol{\beta}$ in (5.6) is obtained by pooling these L estimators, which leads to

$$\widehat{\boldsymbol{\beta}}_{s,t} \quad := \quad (\widehat{\beta}_{s,t,0}, \widehat{\beta}_{s,t,1})^\top = \sum_{\ell=1}^{L} \frac{n_\ell}{n} \widehat{\boldsymbol{\beta}}_\ell, \tag{5.17}$$

with n_ℓ the number of sampled units in the stratum S_ℓ. Note that if the quantiles are used as cut-off points, we have (up to rounding) $n_\ell = n/L$, and $\widehat{\boldsymbol{\beta}}_{s,t}$ is the simple mean of the estimators $\widehat{\boldsymbol{\beta}}_\ell,\ \ell = 1, \ldots, L$.

This leads to the second estimator

$$\widehat{\mu}_{s,t}(t) \quad = \quad \widehat{\beta}_{s,t,0} + \widehat{\beta}_{s,t,1}\, t \tag{5.18}$$

for the dose-response function. Again, ordinary least squares may be replaced by a generalized linear model and appropriate link function to fit other types of outcomes.

5.1.3 Closed form variance estimators

Our objective was to develop closed-form variance estimators for the estimators of the dose-response function presented in equations (5.8) and (5.18). Without loss of generality, we focus on variance estimation for the estimated coefficients of regression $\widehat{\boldsymbol{\beta}}_w$ and $\widehat{\boldsymbol{\beta}}_{s,t}$.

We follow the influence function linearization technique developed by Deville [40] see also Hajage *et al.* [54] For an estimator $\widehat{\boldsymbol{\beta}}$, this technique consists in finding a so-called estimated linearized variable $\widehat{I}_i$, summarizing the variability in the estimation of the parameter. Ideally, the linearized variable should account for all the estimation steps which lead to the estimator $\widehat{\boldsymbol{\beta}}$.

The proposed variance estimator for the weighted estimator $\widehat{\boldsymbol{\beta}}_w$ presented in Section 5.1.2 is given in Section 5.1.3. The proposed variance estimator for the stratified estimator $\widehat{\boldsymbol{\beta}}_{s,t}$ presented in Section 5.1.2 is given in Section 5.1.3.

Weighted treatment effect estimator

The variance estimator for $\widehat{\boldsymbol{\beta}}_w$ is obtained by observing that the coefficient of regression is estimated in a two-step process, involving two estimating equations. First, the unknown parameter $\boldsymbol{\gamma}$ used to compute the weights is obtained by solving the system of estimating equations

$$F_n(\boldsymbol{\gamma}) := \frac{1}{n}\sum_{i=1}^{n} F_i(\boldsymbol{\gamma}) \quad = \quad 0, \tag{5.19}$$

where

$$F_i(\boldsymbol{\gamma}) \quad = \quad \begin{pmatrix} T_i - \mu_T \\ (T_i - \mu_T)^2 - \frac{n-1}{n}\sigma_T^2 \\ (T_i - \tilde{\boldsymbol{Z}}_i^\top \boldsymbol{\alpha})\tilde{\boldsymbol{Z}}_i \\ (T_i - \tilde{\boldsymbol{Z}}_i^\top \boldsymbol{\alpha})^2 - \frac{n-K-1}{n}\sigma^2 \end{pmatrix}. \tag{5.20}$$

Then, the estimator $\widehat{\boldsymbol{\beta}}_w$ is obtained as the solution of the estimating equation

$$H_n(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\beta}) := \frac{1}{n}\sum_{i=1}^{n} w_i(\widehat{\boldsymbol{\gamma}}) H_i(\boldsymbol{\beta}) = 0 \quad \text{with} \quad H_i(\boldsymbol{\beta}) = \tilde{T}_i(Y_i - \tilde{T}_i^\top \boldsymbol{\beta}). \tag{5.21}$$

After some algebra, this leads to the following linearized variable for $\widehat{\boldsymbol{\beta}}_w$:

$$\widehat{I}_{1,i} \quad = \quad \widehat{A}^{-1}\left\{ w_i(\widehat{\boldsymbol{\gamma}}) H_i(\widehat{\boldsymbol{\beta}}) + \widehat{B}\widehat{C}^{-1} F_i(\widehat{\boldsymbol{\gamma}}) \right\}, \tag{5.22}$$

with

$$\begin{aligned}
\widehat{A} &= \frac{1}{n}\sum_{i=1}^{n} w_i(\widehat{\boldsymbol{\gamma}})\tilde{T}_i\tilde{T}_i^\top, \\
\widehat{B} &= \frac{1}{n}\sum_{i=1}^{n} H_i(\widehat{\boldsymbol{\beta}})\nabla w_i(\widehat{\boldsymbol{\gamma}})^\top, \\
\widehat{C} &= \begin{pmatrix} 1 & 0 & \mathbf{0}_{1,K} & 0 \\ 0 & \frac{n-1}{n} & \mathbf{0}_{1,K} & 0 \\ \mathbf{0}_{K1} & \mathbf{0}_{K,1} & \frac{1}{n}\sum_{i=1}^{n}\tilde{\boldsymbol{Z}}_i\tilde{\boldsymbol{Z}}_i^\top & \mathbf{0}_{K,1} \\ 0 & 0 & 0 & \frac{n-K-1}{n} \end{pmatrix},
\end{aligned} \tag{5.23}$$

where ∇ is the differential operator and where $\mathbf{0}_{\bullet,\diamond}$ stands for the null matrix with $\bullet$ rows and $\diamond$ columns.

The resulting variance estimator is

$$\widehat{\mathbb{V}}_{lin}(\widehat{\boldsymbol{\beta}}_w) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\widehat{I}_{1,i}-\bar{I}_1)(\widehat{I}_{1,i}-\bar{I}_1)^\top \quad \text{with} \quad \bar{I}_1 = \frac{1}{n}\sum_{i=1}^{n}\widehat{I}_{1,i}. \tag{5.24}$$

Stratified treatment effect estimator

Inside each stratum $\ell = 1,\ldots,L$, the intermediary estimators $\widehat{\boldsymbol{\beta}}_\ell$ are estimated by solving the estimating equations

$$\Phi(\boldsymbol{\beta}_\ell) := \frac{1}{n_\ell}\sum_{i\in S_\ell}\Phi(Y_i,T_i,\boldsymbol{\beta}_\ell) = 0, \tag{5.25}$$

with

$$\Phi(Y_i,T_i;\boldsymbol{\beta}_\ell) = \begin{pmatrix} Y_i - \tilde{T}_i^\top\boldsymbol{\beta}_\ell \\ T_i(Y_i-\tilde{T}_i^\top\boldsymbol{\beta}_\ell) \end{pmatrix}. \tag{5.26}$$

After some algebra, the linearized variable of $\widehat{\boldsymbol{\beta}}_\ell$ is

$$\widehat{I}_{2,\ell,i} = \frac{1}{\widehat{\sigma}^2_{\ell,T}}\begin{pmatrix} \widehat{\sigma}^2_{\ell,T}+\widehat{m}^2_{\ell,T} & -\widehat{m}_{\ell,T} \\ -\widehat{m}_{\ell,T} & 1 \end{pmatrix} \times \begin{pmatrix} Y_i - \tilde{T}_i^\top\widehat{\boldsymbol{\beta}}_\ell \\ T_i(Y_i-\tilde{T}_i^\top\widehat{\boldsymbol{\beta}}_\ell) \end{pmatrix}, \tag{5.27}$$

where

$$\widehat{m}_{\ell,T} = \frac{1}{n_\ell} \sum_{i \in S_\ell} T_i \quad \text{and} \quad \widehat{\sigma}^2_{\ell,T} = \frac{1}{n_\ell - 1} \sum_{i \in S_\ell} (T_i - \widehat{m}_{\ell,T})^2. \tag{5.28}$$

This leads to the pooled variance estimator

$$\begin{aligned} \widehat{\mathbb{V}}(\widehat{\boldsymbol{\beta}}_{s,t}) &= \sum_{\ell=1}^{L} \frac{1}{n_\ell(n_\ell - 1)} \frac{n_\ell^2}{n^2} \sum_{i \in S_\ell} (\widehat{I}_{2,\ell,i} - \bar{I}_{2,\ell})(\widehat{I}_{2,\ell,i} - \bar{I}_{2,\ell})^\top, \\ \text{with } \bar{I}_{2,\ell} &= \frac{1}{n_\ell} \sum_{i \in S_\ell} \widehat{I}_{2,\ell,i}. \end{aligned} \tag{5.29}$$

5.1.4 Conclusion

Proposed method. In this work, we have proposed closed-form variance estimators for the estimators of the effect of continuous exposure on continuous outcome variables, using the influence function linearization technique. The DRF functions were estimated by weighting the inverse of the GPS or using stratification.

Simulation results. Despite the use of stabilized weights, the variability of the weighted estimator of the DRF was particularly high, and none of the variance estimators (a bootstrap-based estimator, a closed-form estimator especially developed to take into account the estimation step of the GPS, and a sandwich estimator) were able to adequately capture this variability, resulting in coverages below the nominal value, particularly when the proportion of the variation in the quantitative exposure explained by the covariates was large. The stratified estimator was more stable, and variance estimators (a bootstrap-based estimator, a pooled linearized estimator, and a pooled model-based estimator) were more efficient at capturing the empirical variability of the parameters of the DRF. The pooled variance estimators tended to overestimate the variance, whereas the bootstrap estimator, which intrinsically takes into account the estimation step of the GPS, resulted in correct variance estimations and coverage rates.

5.2 S-estimation in linear models with structured covariance matrices

Motivations :

- Linear mixed models are widely used and provide an approach for analyzing correlated responses, such as longitudinal data, growth data, or repeated measurements.
- To be resistant against outliers, S-estimators have been investigated.

Contributions of the Chapter [VG12] :

- Providing a unified approach to S-estimation in balanced linear models with structured covariance matrices
- Providing sufficient conditions for the existence of S-functionals and S-estimators, establish their asymptotic properties, such as consistency and asymptotic normality, and derive their robustness properties in terms of breakdown point and influence function.

Collaborations :

- *Anne Ruiz-Gazen* (Toulouse school of economics, Toulouse) and *Rik Lopuhaä* (Delft University of Technology)

Perspectives :

- S-estimation in non-balanced linear mixed models

5.2.1 Introduction

Problem and objective

Problem. Linear mixed-effects models are widely used for the analysis of correlated responses, such as longitudinal data, growth data, or repeated measurements. However, they are sensitive to extreme and/or outlier values.

Objective. Therefore, it can be valuable to propose robust estimators for this type of model.

For the sake of compactness I will only introduce the main definitions and models in this manuscript. All the detailed properties, proofs, or examples can be found in the dedicated paper [VG12].

Literature

In linear models, each subject i, $i = 1, \ldots, n$, is observed at k_i occasions, and the vector of responses $\mathbf{y}_i$ is assumed to arise from the model

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{u}_i,$$

where $\mathbf{X}_i$ is the design matrix for the ith subject and $\mathbf{u}_i$ is a vector whose covariance matrix can be used to model the correlation between the responses. One possibility is the linear mixed effects model, in which the random effects together with the measurement error yield a specific covariance structure depending on a vector $\boldsymbol{\theta}$ consisting of some unknown covariance parameters. Other covariance structures may arise, for example, if the $\mathbf{u}_i$ are the outcome of a time series, see e.g., [67] or [49], for different possible covariance structures.

Maximum likelihood estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ has been studied, e.g., in [57, 74], see also [49, 36]. To be resistant against outliers, robust methods have been investigated for linear mixed effects models, e.g., in [92, 25, 24, 59, 2, 18]. This mostly concerns S-estimators, originally introduced in the multiple regression context by Rousseeuw and Yohai [98] and extended to multivariate location and scatter in [33, 79], to multivariate regression in [108], and to linear mixed effects models in [25, 59, 18]. S-estimators are well known smooth versions of the minimum volume ellipsoid estimator [97] that are highly resistant against outliers. As such, S-estimators have gained popularity as robust estimators, but they may also serve as initial estimators to further improve efficiency. However, the theory about these estimators is far from complete, even in balanced models where the number of observed responses is the same for all subjects.

5.2.2 Balanced models with structured covariances

We consider independent observations $(\mathbf{y}_1, \mathbf{X}_1), \ldots, (\mathbf{y}_n, \mathbf{X}_n)$, for which we assume the following model

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{u}_i, \quad i = 1, \ldots, n, \tag{5.30}$$

where $\mathbf{y}_i \in \mathbb{R}^k$ contains repeated measurements for the i-th subject, $\boldsymbol{\beta} \in \mathbb{R}^q$ is an unknown parameter vector, $\mathbf{X}_i \in \mathbb{R}^{k\times q}$ is a known design matrix, and $\mathbf{u}_i \in \mathbb{R}^k$ are unobservable independent mean zero random vectors with covariance matrix $\mathbf{V} \in \mathrm{PDS}(k)$, the class

of positive definite symmetric $k \times k$ matrices. The model is balanced in the sense that all $\mathbf{y}_i$ have the same dimension. Furthermore, we consider a structured covariance matrix, that is, the matrix $\mathbf{V} = \mathbf{V}(\boldsymbol{\theta})$ is a known function of unknown covariance parameters combined in a vector $\boldsymbol{\theta} \in \mathbb{R}^l$. Table 5.1 refers to different structured covariance matrices. An important case of interest is the (balanced) linear mixed effects model

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}\gamma_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n.$$

This model arises from $\mathbf{u}_i = \mathbf{Z}\boldsymbol{\Gamma}_i + \boldsymbol{\epsilon}_i$, for $i = 1, \ldots, n$, where $\mathbf{Z} \in \mathbb{R}^{k\times g}$ is known and $\boldsymbol{\Gamma}_i \in \mathbb{R}^g$ and $\boldsymbol{\epsilon}_i \in \mathbb{R}^k$ are independent mean zero random variables, with unknown covariance matrices $\mathbf{G}$ and $\mathbf{R}$, respectively. In this case $\mathbf{V}(\boldsymbol{\theta}) = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R}$ and $\boldsymbol{\theta} = (\text{vech}(\mathbf{G})^T, \text{vech}(\mathbf{R})^T)^T$, where

$$\text{vech}(\mathbf{A}) = (a_{1,1}, \ldots, a_{k,1}, a_{2,2}, \ldots, a_{k,k}) \tag{5.31}$$

is the unique $k(k+1)/2$-vector that stacks the columns of the lower triangle elements of a symmetric matrix $\mathbf{A}$. In full generality, the model is usually overparametrized and one may run into identifiability problems. A more feasible example is obtained by taking $\mathbf{R} = \sigma_0^2\mathbf{I}_k$, $\mathbf{Z} = [\mathbf{Z}_1 \cdots \mathbf{Z}_r]$ and $\boldsymbol{\Gamma}_i = (\gamma_{i,1} \ldots, \gamma_{i,r})^T$, where the $\mathbf{Z}_j$'s are known $k \times g_j$ design matrices and the $\gamma_{i,j} \in \mathbb{R}^{g_j}$ are independent mean zero random variables with covariance matrix $\sigma_j^2\mathbf{I}_{g_j}$, for $j = 1, \ldots, r$. This leads to

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \sum_{j=1}^{r} \mathbf{Z}_j\gamma_{i,j} + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n, \tag{5.32}$$

with $\mathbf{V}(\boldsymbol{\theta}) = \sum_{j=1}^r \sigma_j^2\mathbf{Z}_j\mathbf{Z}_j^T + \sigma_0^2\mathbf{I}_k$ and $\boldsymbol{\theta} = (\sigma_0^2, \sigma_1^2, \ldots, \sigma_r^2)$.

In this work, we assume that the parameter $\boldsymbol{\theta}$ is identifiable in the sense that,

$$\mathbf{V}(\boldsymbol{\theta}_1) = \mathbf{V}(\boldsymbol{\theta}_2) \quad \Rightarrow \quad \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2. \tag{5.33}$$

This may not be true in general for the linear mixed effects model. For linear mixed effects models in (5.32), identifiability of $\boldsymbol{\theta} = (\sigma_0^2, \sigma_1^2, \ldots, \sigma_r^2)$ holds for particular choices of the design matrices $\mathbf{Z}_1, \ldots, \mathbf{Z}_r$.

TABLE 5.1 – Structured covariances.

Structure	l	Description
Independent observations	1	$\mathbf{V} = \sigma^2 \mathbf{I}_k$
Compound symmetry	2	$\mathbf{V} = \sigma_b^2 \mathbf{1}_k \mathbf{1}_k^T + \sigma^2 \mathbf{I}_k$
Linear mixed effects	$\frac{1}{2}g(g+1)+1$	$\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \sigma^2 \mathbf{I}_k$ $\mathbf{Z}$ known matrix
Factor analytic	$kg - \frac{1}{2}g(g-1)+k$	$\mathbf{V} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi}$ $\Lambda = k \times g$ matrix; Ψ diagonal
First order autoregressive	2	$v_{i,j} = \sigma^2 \rho^{\lvert i-j \rvert}$
Banded, or general autoregressive	k	$v_{i,j} = \theta_{\lvert i-j \rvert + 1}$
Unstructured	$\frac{1}{2}k(k+1)$	$v_{1,1} = \theta_1, v_{1,2} = \theta_2, \ldots, v_{k,k} = \theta_l$.

5.2.3 Definitions

We start by representing our observations as points in $\mathbb{R}^k \times \mathbb{R}^{kq}$ in the following way. For $r = 1, \ldots, k$, let $\mathbf{x}_r^T$ denote the r-th row of the $k \times q$ matrix $\mathbf{X}$, so that $\mathbf{x}_r \in \mathbb{R}^q$. We represent the pair $\mathbf{s} = (\mathbf{y}, \mathbf{X})$ as an element in $\mathbb{R}^k \times \mathbb{R}^{kq}$ defined by $\mathbf{s}^T = (\mathbf{y}^T, \mathbf{x}_1^T, \ldots, \mathbf{x}_k^T)$. In this way our observations can be represented as $\mathbf{s}_1, \ldots, \mathbf{s}_n$, with $\mathbf{s}_i = (\mathbf{y}_i, \mathbf{X}_i) \in \mathbb{R}^k \times \mathbb{R}^{kq}$.

S-estimator

S-estimators are defined by means of a function $\rho : \mathbb{R} \to [0, \infty)$ that satisfies the following properties

(R1) ρ is symmetric around zero with $\rho(0) = 0$ and ρ is continuous at zero;

(R2) There exists a finite constant $c_0 > 0$, such that ρ is non-decreasing on $[0, c_0]$ and constant on $[c_0, \infty)$; put $a_0 = \sup \rho$.

The S-estimator $\boldsymbol{\xi}_n = (\boldsymbol{\beta}_n, \boldsymbol{\theta}_n)$ is defined as the solution to the following minimization problem

$$\begin{aligned} &\min_{\boldsymbol{\beta},\boldsymbol{\theta}} \det(\mathbf{V}(\boldsymbol{\theta})) \\ &\text{subject to} \\ &\frac{1}{n}\sum_{i=1}^{n} \rho\left(\sqrt{(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^T \mathbf{V}(\boldsymbol{\theta})^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})}\right) \leq b_0, \end{aligned} \tag{5.34}$$

where the minimum is taken over all $\boldsymbol{\beta} \in \mathbb{R}^q$ and $\boldsymbol{\theta} \in \mathbb{R}^l$, such that $\mathbf{V}(\boldsymbol{\theta}) \in \text{PDS}(k)$, with ρ satisfying (R1)-(R2).

The S-estimator defined by (5.34) for the setup in (5.30) includes several specific cases that have been considered in the literature. Copt and Victoria-Feser [25] and Chervoneva and Vishnyakov [18] consider S-estimators for the parameters in the linear mixed effects model (5.32).

The constant $0 < b_0 < a_0$ in (5.34) can be chosen in agreement with an assumed underlying distribution. For the multivariate regression model in [108], it is assumed that $\mathbf{y}_i \mid \mathbf{X}_i$ has an elliptically contoured density of the form

$$f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\mathbf{y}) = \det(\boldsymbol{\Sigma})^{-1/2} h\left((\mathbf{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})\right), \tag{5.35}$$

with $\boldsymbol{\mu} = \mathbf{X}_i\boldsymbol{\beta}$ and $\boldsymbol{\Sigma} = \mathbf{V}(\boldsymbol{\theta})$ and $h : [0, \infty) \to [0, \infty)$. For the linear mixed effects model in [25], it is assumed that $\mathbf{y}_i \mid \mathbf{X}_i$ has a multivariate normal distribution, which is a special case of (5.35) with $h(t) = (2\pi)^{-k/2} \exp(-t/2)$. When the underlying distribution corresponds to a density of the form (5.35), then a natural choice is $b_0 = \mathbb{E}_{\mathbf{0},\mathbf{I}}\rho(\|\mathbf{z}\|)$, where $\mathbf{z}$ has density (5.35) with $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\mathbf{0}, \mathbf{I}_k)$. Finally, it should be emphasized that the ratio b_0/a_0 determines the breakdown point of the S-estimator (see Theorem 6.1 in [VG12]), as well as its limiting variance (see Corollary 9.2 in [VG12]). By choosing the constant c_0 in (R2) one then has to make a trade-off between robustness and efficiency.

Remark 5.1. *Clearly, the definition of the S-estimator in* (5.34) *has great similarities with the S-estimator for multivariate location and covariance (see [32] and [79]), defined*

as the solution $(\mathbf{t}_n, \mathbf{C}_n)$ *to the minimization problem*

$$\begin{aligned} &\min_{\mathbf{t},\mathbf{C}} \det(\mathbf{C}) \\ &\textit{subject to} \\ &\frac{1}{n}\sum_{i=1}^{n} \rho\left(\sqrt{(\mathbf{y}_i - \mathbf{t})^T \mathbf{C}^{-1} (\mathbf{y}_i - \mathbf{t})}\right) \leq b_0, \end{aligned} \tag{5.36}$$

where the minimum is taken over all $\mathbf{t} \in \mathbb{R}^k$ *and* $\mathbf{C} \in PDS(k)$*. Even more so, if all* $\mathbf{X}_i$ *are assumed to be equal to the same design matrix* $\mathbf{X}$ *of full rank, as was done in [25, 24]. However, there is a subtle, but important difference between minimization problems* (5.36) *and* (5.34)*. The important difference is that in* (5.36) *we minimize over* all *positive definite symmetric* $k \times k$ *matrices* $\mathbf{C}$*, whereas in* (5.34)*, we* only *minimize over positive definite symmetric* $k \times k$ *matrices* $\mathbf{V}(\boldsymbol{\theta})$*, which can arise as the image of the mapping* $\boldsymbol{\theta} \mapsto \mathbf{V}(\boldsymbol{\theta})$*. The latter collection is a subset of the other :*

$$\left\{\mathbf{V}(\boldsymbol{\theta}) \in \mathrm{PDS}(k) : \boldsymbol{\theta} \in \mathbb{R}^l\right\} \subset PDS(k),$$

and will typically be a strictly smaller subset. This means that the properties of $\mathbf{V}(\boldsymbol{\theta}_n)$ *and* $\mathbf{C}_n$ *are related, but the properties of* $\mathbf{V}(\boldsymbol{\theta}_n)$ cannot *simply be derived from properties of* $\mathbf{C}_n$*, not even in the case where all* $\mathbf{X}_i$ *are equal to the same* $\mathbf{X}$*. In fact, this will lead to limiting covariances that differ from the ones found in [25].*

S-functional

The concept of S-functional is needed to investigate the local robustness properties of the corresponding S-estimator, such as the influence function. Let $\mathbf{s} = (\mathbf{y}, \mathbf{X})$ have a probability distribution P on $\mathbb{R}^k \times \mathbb{R}^{kq}$. The S-functional $\boldsymbol{\xi}(P) = (\boldsymbol{\beta}(P), \boldsymbol{\theta}(P))$ is defined as the solution to the following minimization problem :

$$\begin{aligned} &\min_{\boldsymbol{\beta},\boldsymbol{\theta}} \det(\mathbf{V}(\boldsymbol{\theta})) \\ &\text{subject to} \\ &\int \rho\left(\sqrt{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}(\boldsymbol{\theta})^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}\right) \mathrm{d}P(\mathbf{y}, \mathbf{X}) \leq b_0, \end{aligned} \tag{5.37}$$

where the minimum is taken over all $\boldsymbol{\beta} \in \mathbb{R}^q$ and $\boldsymbol{\theta} \in \mathbb{R}^l$, such that $\mathbf{V}(\boldsymbol{\theta}) \in \mathrm{PDS}(k)$, with ρ satisfying (R1)-(R2).

As a special case, we obtain the S-estimator $\boldsymbol{\xi}_n = (\boldsymbol{\beta}_n, \boldsymbol{\theta}_n)$ by taking $P = \mathbb{P}_n$, the empirical measure corresponding to the observations $(\mathbf{y}_1, \mathbf{X}_1), \ldots, (\mathbf{y}_n, \mathbf{X}_n)$. In view of this connection, the existence and consistency of solutions of (5.34) will follow from general results on the existence and the continuity of solutions to (5.37).

5.2.4 Theoretical properties

We established the robustness properties of the S-estimators in terms of breakdown point and influence function. Let's recall their definitions.

Breakdown point. Consider a collection of points $\mathcal{S}_n = \{\mathbf{s}_i = (\mathbf{y}_i, \mathbf{X}_i), i = 1, \ldots, n\} \subset \mathbb{R}^k \times \mathcal{X}$. To emphasize the dependence on the collection $\mathcal{S}_n$, denote by $\boldsymbol{\xi}_n(\mathcal{S}_n) = (\boldsymbol{\beta}_n(\mathcal{S}_n), \boldsymbol{\theta}_n(\mathcal{S}_n))$, the S-estimator, as defined in (5.34). To investigate the global robustness of S-estimators, we compute that finite-sample (replacement) breakdown point. For a given collection $\mathcal{S}_n$ the finite-sample breakdown point (see Donoho and Huber [41]) of a regression S-estimator $\boldsymbol{\beta}_n$ is defined as the smallest proportion of points from $\mathcal{S}_n$ that one needs to replace in order to carry the estimator over all bounds. More precisely,

$$\epsilon_n^*(\boldsymbol{\beta}_n, \mathcal{S}_n) = \min_{1 \le m \le n} \left\{ \frac{m}{n} : \sup_{\mathcal{S}'_m} \|\boldsymbol{\beta}_n(\mathcal{S}_n) - \boldsymbol{\beta}_n(\mathcal{S}'_m)\| = \infty \right\}, \tag{5.38}$$

where the minimum runs over all possible collections $\mathcal{S}'_m$ that can be obtained from $\mathcal{S}_n$ by replacing m points of $\mathcal{S}_n$ by arbitrary points in $\mathbb{R}^k \times \mathcal{X}$.

For any $k \times k$ matrix $\mathbf{A}$, let $\lambda_k(\mathbf{A}) \le \cdots \le \lambda_1(\mathbf{A})$ denote the eigenvalues of $\mathbf{A}$. The estimator $\boldsymbol{\theta}_n$ determines the covariance estimator $\mathbf{V}_n = \mathbf{V}(\boldsymbol{\theta}_n)$. For this reason, it seems natural to let the breakdown point of $\boldsymbol{\theta}_n$ correspond to the breakdown of a covariance estimator. We define the finite sample (replacement) breakdown point of the S-estimator $\boldsymbol{\theta}_n$ at a collection $\mathcal{S}_n$, as

$$\epsilon_n^*(\boldsymbol{\theta}_n, \mathcal{S}_n) = \min_{1 \le m \le n} \left\{ \frac{m}{n} : \sup_{\mathcal{S}'_m} \operatorname{dist}(\mathbf{V}(\boldsymbol{\theta}_n(\mathcal{S}_n))), \mathbf{V}(\boldsymbol{\theta}_n(\mathcal{S}'_m)) = \infty \right\}, \tag{5.39}$$

with $\operatorname{dist}(\cdot, \cdot)$ defined as $\operatorname{dist}(\mathbf{A}, \mathbf{B}) = \max\left\{ |\lambda_1(\mathbf{A}) - \lambda_1(\mathbf{B})|, |\lambda_k(\mathbf{A})^{-1} - \lambda_k(\mathbf{B})^{-1}| \right\}$, where

the minimum runs over all possible collections $\mathcal{S}'_m$ that can be obtained from $\mathcal{S}_n$ by replacing m points of $\mathcal{S}_n$ by arbitrary points in $\mathbb{R}^k \times \mathcal{X}$. So the breakdown point of $\boldsymbol{\theta}_n$ is the smallest proportion of points from $\mathcal{S}_n$ that one needs to replace in order to make the largest eigenvalue of $\mathbf{V}(\boldsymbol{\theta}(\mathcal{S}'_m))$ arbitrarily large (explosion), or to make the smallest eigenvalue of $\mathbf{V}(\boldsymbol{\theta}(\mathcal{S}'_m))$ arbitrarily small (implosion).

Influence function. For $0 < h < 1$ and $\mathbf{s} = (\mathbf{y}, \mathbf{X}) \in \mathbb{R}^k \times \mathbb{R}^{kq}$ fixed, define the perturbed probability measure $P_{h,\mathbf{s}} = (1-h)P + h\delta_{\mathbf{s}}$, where $\delta_{\mathbf{s}}$ denotes the Dirac measure at $\mathbf{s} \in \mathbb{R}^k \times \mathbb{R}^{kq}$. The *influence function* of the functional $\boldsymbol{\xi}(\cdot)$ at probability measure P, is defined as

$$\mathrm{IF}(\mathbf{s}; \boldsymbol{\xi}, P) = \lim_{h \downarrow 0} \frac{\boldsymbol{\xi}((1-h)P + h\delta_{\mathbf{s}}) - \boldsymbol{\xi}(P)}{h}, \tag{5.40}$$

if this limit exists. In contrast to the global robustness measured by the breakdown point, the influence function measures the local robustness. It describes the effect of an infinitesimal contamination at a single point $\mathbf{s}$ on the functional (see Hampel [56]). Good local robustness is therefore illustrated by a bounded influence function.

We provide sufficient conditions for the existence of S-functionals and S-estimators, establish their asymptotic properties, such as consistency and asymptotic normality, and derive their robustness properties in terms of breakdown point and influence function. These conditions concern the ρ function, the distribution of the observations, and the covariance $\mathbf{V}(\boldsymbol{\theta})$. All results are obtained for a large class of identifiable covariance structures and are established under very mild conditions on the distribution of the observations, which goes far beyond models with elliptically contoured densities.

5.2.5 Conclusion

Proposed method. We proposed a unified approach to S-estimation in balanced linear mixed-effects models with structured covariance matrices. Our primary focus is on S-estimators for linear mixed-effects models, but our approach also encompasses S-estimators in several other standard multivariate models, such as multiple regression, multivariate regression, and multivariate location and scatter. We illustrate our findings through an application of data from a clinical trial on the treatment of children exposed to lead.

5.3 Perspectives

5.3.1 Analysis of longitudinal data with outliers

In our current work, we focus on balanced linear mixed effects models and provide an overview of robust estimation methods that have been studied and revisited in recent years, such as the S, the MM, and the tau estimators, and their composite counterparts. We compare them in different contamination models. In the robust statistical literature, there exist two contamination models : the Classical (Tukey-Huber) Contamination Model (CCM) and the Independent Contamination Model (ICM). While CCM (also called “case-wise" contamination) considers that the contamination is at the level of the subjects or cases, ICM (also called "cell-wise" contamination) considers the possibility to contaminate data sets at the level of the cells. Initially, the robust statistics literature focused on the CCM context and proposed robust estimators that were able to cope with a proportion of outliers close to 50% without breaking down. However, such estimators may not be robust in the ICM context, since even a small fraction of contaminated cells may lead to more than 50% of contaminated cases. We will apply these methods to clinical data in collaboration with Benoit Lepage (INSERM UMR 1295 CERPOP).

In practice, the number of repeated measurements of the variable of interest may differ according to each individual. Indeed, some individuals miss visits or drop out of the study early. In this case, the mixed model is said to be unbalanced. We will therefore propose a robust estimator in the case of the unbalanced mixed model.

5.3.2 Analysis of functional data

I am interested in functional data modeling within the framework of a real industrial application, in collaboration with Madison Giacofci and Valérie Monbet (section 5.3.2). Our project is to propose a PhD topic in 2025 on the modeling of environmental contamination data (spectrometry data) from non-targeted analysis methods in collaboration with researchers at the institute IRSET. Indeed, spectrometry data can be modeled using a functional approach (section 5.3.2).

Functional data with grouped structures

Objective. Motivated by a real data application for industry, the objective is the prediction and explanation of a force curve using the geometric characteristics of a rubber joint. The relationship between the functional responses (force) and the geometric characteristics is not homogeneous for all rubber joints.

Proposed method. We propose a mixture model of canonical correlation analysis. The responses observed are functional data that are characterized by explanatory variables. The infinite-dimension nature of the data is accounted through a functional canonical correlation analysis. The relationship between functional responses and explanatory variables being not homogeneous for all individuals, we consider a mixture of regression models. The originality of this work lies in the simultaneous clustering and prediction of the individual canonical component scores. This is achieved by adopting a probabilistic approach of the canonical correlation analysis. Estimation of the parameters of the models is driven by the EM algorithm where both cluster labels and individual functional canonical component analysis scores are latent.

Modeling environmental contamination data from non-targeted analysis methods

Problem. Non-targeted analyses based on the use of liquid chromatography coupled with high-resolution mass spectrometry (LC-HRMS) offer the promise of globally identifying and even quantifying pollutants present in biological matrices such as urine, blood, and hair [12, 31]. The mass spectrometer acts as a detector, measuring the mass-to-charge ratio (m/z) of ions detected in a sample, as well as the associated abundance. Upstream liquid chromatography allows the separation of compounds to simplify a complex sample. This results in three-dimensional data forming peaks (m/z, intensity, retention time). In a non-targeted approach, we do not focus on predefined specific pollutants but rather on the entire chemical fingerprint characterized by multiple peaks corresponding to identified or unidentified molecules. Several challenges remain to be addressed in order to effectively exploit this massive data : the pollutants of interest are often of low abundance and are masked by endogenous compounds, making them particularly difficult to detect. Moreover, not all peaks can be described by the same "mathematical curve" (i.e., Gaussian). Finally, the techniques used to record these data are specific to each laboratory, and the

joint analysis of exposure profiles produced by different laboratories remains an unresolved challenge.

Objective. We may deal with two objectives : the first is to relate this curve with a health event (or a lifetime) to identify the associated peaks and then interpret them in terms of molecules in a second time trying to annotate them. The second is to identify homogeneous exposure profiles.

Project.

Existing Approach. This overall objective is currently addressed in two main steps in the literature. The first step, a preprocessing phase carried out simultaneously with the acquisition of the spectra, consists of reducing the entire spectrum to a position/intensity matrix that summarizes the molecular information of the sample. This matrix is then used, in a second step, as input for classical learning models, either supervised or unsupervised, to explain/predict an event or identify individual profiles. Such an approach has several limitations. First, the preprocessing of these spectra through these methods involves multiple steps [96]. These different steps depend on numerous parameters that need to be specified, thereby increasing user-related subjectivity. One of the challenges is, therefore, to reduce the number of parameters or automate their selection. Moreover, each step introduces statistical errors that are rarely quantified or considered in existing methods. It is thus necessary to quantify the uncertainty arising from each step of the processing as a way to ensure a better evaluation of data quality.

Proposed method. Our project aims to adopt a more global approach to reduce preprocessing steps and the uncertainty arising from errors propagated through successive steps [103]. To achieve this, we will propose a functional modeling of the spectrum using flexible function bases adapted to the characteristics of the acquired spectra. One of the challenges with spectra is that the peaks observed in LC-HRMS data for different individuals are not properly aligned ; we will integrate an alignment step into our models based on optimal transport and the Wasserstein distance [105]. Moreover, pollutants present in biological samples typically correspond to small peaks whose intensity is close to the noise level. Our model will thus need to account for this in order to distinguish peaks associated with real molecules from those corresponding to noise. Additionally, we will

take into account various sources of variability, such as those due to different laboratory techniques or group structures, in the final model, using mixed effects. We will also define a penalty term specifically suited for the selection of curve segments. This modeling will allow us to identify, without prior assumptions, the pollutants with the most significant effect on a health event.

BIBLIOGRAPHY

[1] A. Abadie et G. W. Imbens, « Matching on the Estimated Propensity Score », English, in : *Econometrica* 84.*2* (mars 2016), p. 781-807, issn : 1468-0262.

[2] C. Agostinelli et V. Yohai, « Composite robust estimators for linear mixed models », in : *Journal of the American Statistical Association* 111.*516* (2016), p. 1764-1774.

[3] M. Aitkin, « A Note on the Regression Analysis of Censored Data », English, in : *Technometrics* 23.*2* (1981), p. 161-163, issn : 00401706.

[4] P. K Andersen et R. D. Gill, « Cox's Regression Model for Counting Processes : A Large Sample Study », in : *The Annals of Statistics* 10.*4* (1982), p. 1100-1120, issn : 00905364.

[5] P. C. Austin, « Assessing the performance of the generalized propensity score for estimating the effect of quantitative or continuous exposures on binary outcomes », in : *Statistics in medicine* 37.*11* (2018), p. 1874-1894.

[6] D.J. Bartholomew, M. Knott et I. Moustaki, *Latent Variable Models and Factor Analysis : A Unified Approach*, (3rd ed). Wiley, 2011.

[7] T. R. Belin et D. B Rubin, « A Method for Calibrating False-Match Rates in Record Linkage », in : *Journal of the American Statistical Association* 90.*430* (1995), p. 694-707.

[8] J. Bezin et al., « The national healthcare system claims databases in France, SNIIRAM and EGB : Powerful tools for pharmacoepidemiology », in : *Pharmacoepidemiology and Drug Safety* 26.*8* (2017), p. 954-962.

[9] M. Bia et A. Mattei, « A Stata package for the estimation of the dose-response function through adjustment for the generalized propensity score », in : *The Stata Journal* 8.*3* (2008), p. 354-373.

[10] J. F. Bobb et R. Varadhan, *turboEM : A Suite of Convergence Acceleration Schemes for EM, MM and Other Fixed-Point Algorithms*, R package version 2018.1, 2018.

[11] N. Breslow, « Covariance analysis of censored survival data », in : *Biometrics* (1974), p. 89-99.

[12] J. Chaker et al., « Comprehensive evaluation of blood plasma and serum sample preparations for HRMS-based chemical exposomics : overlaps and specificities », in : *Analytical Chemistry* 94.*2* (2022), p. 866-874.

[13] R. Chambers, « Regression Analysis of Probability-Linked Data », in : *Statistics New Zealand* (2009).

[14] R. Chambers et al., « Domain estimation under informative linkage », in : *Statistical Theory and Related Fields* 3.*2* (2019), p. 90-102.

[15] B. E. Chen et al., « Competing risks analysis of correlated failure time data », in : *Biometrics* 64.*1* (2008), p. 172-179.

[16] W.-Y. Chen et al., « Transfer Neural Trees : Semi-Supervised Heterogeneous Domain Adaptation and Beyond », in : *IEEE Transactions on Image Processing* 28.*9* (sept. 2019), p. 4620-4633, issn : 1941-0042.

[17] Y.-H. Chen, « Semiparametric marginal regression analysis for dependent competing risks under an assumed copula », English, in : *Journal of the Royal Statistical Society, Series B (Methodological)* 72 (2010), p. 235-251, issn : 1369-7412.

[18] I. Chervoneva et M. Vishnyakov, « Generalized S-estimators for linear mixed effects models », in : *Statistica Sinica* 24.*3* (2014), p. 1257-1276, issn : 10170405, 19968507.

[19] P. Christen, « Automatic Record Linkage Using Seeded Nearest Neighbour and Support Vector Machine Classification », in : *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, Las Vegas, Nevada, USA : Association for Computing Machinery, 2008, 151–159, isbn : 9781605581934.

[20] P. Christen, *Data Matching : Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*, Springer Publishing Company, Incorporated, 2012.

[21] P. Christen et W. E. Winkler, « Record Linkage », in : *Encyclopedia of Machine Learning and Data Mining*, Boston, MA : Springer US, 2017, p. 1066-1075.

[22] S. R. Cole, H. Chu et S. Greenland, « Maximum Likelihood, Profile Likelihood, and Penalized Likelihood : A Primer », in : *American Journal of Epidemiology* 179.*2* (oct. 2013), p. 252-260, issn : 0002-9262.

[23] J. B. Copas et F. J Hilton, « Record Linkage : Statistical Models for Matching Computer Records », in : *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 153.*3* (1990), p. 287-320, issn : 09641998, 1467985X.

[24] S. Copt et S. Heritier, « Robust alternatives to the F-test in mixed linear models based on MM-estimates », in : *Biometrics* 63.*4* (2007), p. 1045-1052.

[25] S. Copt et M.-P. Victoria-Feser, « High-breakdown inference for mixed linear models », in : *Journal of the American Statistical Association* 101.*473* (2006), p. 292-300.

[26] N. Courty, R. Flamary et D. Tuia, « Domain adaptation with regularized optimal transport », in : *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases 2014*, LNCS, Nancy, France, 2014, p. 1-16.

[27] N. Courty et al., « Joint distribution optimal transportation for domain adaptation », in : *Advances in Neural Information Processing Systems*, 2017, p. 3730-3739.

[28] N. Courty et al., « Optimal transport for domain adaptation », in : *IEEE transactions on pattern analysis and machine intelligence* 39.*9* (2016), p. 1853-1865.

[29] D. R. Cox, « Regression Models and Life-Tables », in : *Journal of the Royal Statistical Society. Series B (Methodological)* 34.*2* (1972), p. 187-220, issn : 00359246.

[30] B. B. Damodaran et al., « DeepJDOT : Deep Joint Distribution Optimal Transport for Unsupervised Domain Adaptation », in : *ECCV 2018 - 15th European Conference on Computer Vision*, t. 11208, LNCS, Munich, Germany : Springer, sept. 2018, p. 467-483.

[31] A. David et al., « Towards a comprehensive characterisation of the human internal chemical exposome : Challenges and perspectives », in : *Environment International* 156 (2021), p. 106630.

[32] P. L. Davies, « Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices », in : *Ann. Statist.* 15.*3* (1987), p. 1269-1292, ISSN : 0090-5364.

[33] P Laurie Davies, « Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices », in : *The Annals of Statistics* (1987), p. 1269-1292.

[34] O. Day et T. M. Khoshgoftaar, « A survey on heterogeneous transfer learning », in : *Journal of Big Data* 4 (2017).

[35] C. Delpierre et al., « What role does socio-economic position play in the link between functional limitations and self-rated health : France vs. USA ? », in : *European Journal of Public Health* 22.*3* (2012), p. 317-321.

[36] E. Demidenko, *Mixed models : theory and applications with R*, John Wiley & Sons, 2013.

[37] A. Dempster, N. Laird et B. D. Rubin, « Maximum Likelihood From Incomplete Data Via The EM algorithm », in : *Journal of the Royal Statistical Society. Series B (Methodological)* 39 (jan. 1977), p. 1-38.

[38] N. Deresa et I. Van Keilegom, « A multivariate normal regression model for survival data subject to different types of dependent censoring », English, in : *Computational Statistics & Data Analysis* 144 (2020), ISSN : 0167-9473.

[39] N. W. Deresa et I. Van Keilegom, « Flexible parametric model for survival data subject to dependent censoring », in : *Biometrical Journal* 62.*1* (2020), p. 136-156, ISSN : 0323-3847.

[40] J.-C Deville, « Variance Estimation for Complex Statistics and Estimators : Linearization and Residual Techniques », in : *Survey methodology* 25.*2* (1999), p. 193-204.

[41] David Donoho et Peter J. Huber, « The notion of breakdown point », in : *A Festschrift for Erich L. Lehmann*, Wadsworth Statist./Probab. Ser. Wadsworth, Belmont, CA, 1983, p. 157-184.

[42] M. D'Orazio, M. Di Zio et M. Scanu, *Statistical matching : Theory and practice*, John Wiley & Sons, 2006.

[43] T. Emura et Y. H. Chen, « Gene selection for survival data under dependent censoring : a copula-based approach. », in : *Statistical Methods in Medical Research* 25.*6* (2016), p. 2840-2857.

[44] T. Emura et al., « Comparison of the marginal hazard model and the subdistribution hazard model for competing risks under an assumed copula », in : *Statistical methods in medical research* 29.*8* (2020), p. 2307-2327.

[45] T. Enamorado, « Active Learning for Probabilistic Record Linkage », in : *Social Science Research Network (SSRN)* (2018).

[46] T. Enamorado, B. Fifield et K. Imai, « Using a probabilistic model to assist merging of large-scale administrative records », in : *American Political Science Review* 113 (2019), p. 353-371.

[47] I. Fellegi et A. Sunter, « A Theory for Record Linkage », in : *Journal of the American Statistical Association* 64 (déc. 1969), p. 1183-1210.

[48] J. P. Fine et R. J. Gray, « A Proportional Hazards Model for the Subdistribution of a Competing Risk », in : *Journal of the American Statistical Association* 94.*446* (1999), p. 496-509.

[49] G.M. Fitzmaurice et N.M. Laird, « Ware JH », in : *Applied Longitudinal Analysis Second Edition ed : Wiley* (2011).

[50] C. A. Flores et al., « Estimating the Effects of Length of Exposure to Instruction in a Training Program : The Case of Job Corps », in : *The Review of Economics and Statistics* 94.*1* (2012), p. 153-171.

[51] S. J. Grannis et al., « Analysis of a Probabilistic Record Linkage Technique without Human Review », in : *AMIA ... Annual Symposium proceedings. AMIA Symposium* (2003), p. 259-63.

[52] P. J. Green, « On Use of the EM Algorithm for Penalized Likelihood Estimation », in : *Journal of the Royal Statistical Society. Series B (Methodological)* 52.*3* (1990), p. 443-452, issn : 00359246.

[53] I. D. Ha et al., « Analysis of clustered competing risks data using subdistribution hazard models with multivariate frailties. », eng, in : *Statistical Methods in Medical Research* 25 (6 2016), p. 2488-2505.

[54] D. Hajage et al., « Closed-form variance estimator for weighted propensity score estimators with survival outcome », in : *Biometrical Journal* 60.*6* (2018), p. 1151-1163.

[55] D. Hajage et al., « On the Use of Propensity Scores in Case of Rare Exposure », English, in : *BMC medical research methodology* 16 (mars 2016), p. 38, issn : 1471-2288.

[56] F. R. Hampel, « The influence curve and its role in robust estimation », in : *J. Amer. Statist. Assoc.* 69 (1974), p. 383-393, issn : 0162-1459.

[57] H. O. Hartley et J. N.K. Rao, « Maximum-likelihood estimation for the mixed analysis of variance model », in : *Biometrika* 54.*1-2* (1967), p. 93-108.

[58] B. Hejblum et al., « Probabilistic record linkage of de-identified research datasets with discrepancies using diagnosis codes », in : *Scientific Data* 6 (2019).

[59] S. Heritier et al., *Robust methods in biostatistics*, John Wiley & Sons, 2009.

[60] T. Herzog, F. Scheuren et W. Winkler, *Data Quality and Record Linkage Techniques*, Springer-Verlag New York, 2007.

[61] K. Hirano et G. Imbens, *The Propensity Score with Continuous Treatments," in A. Gelman and X.-L. Meng (eds.), Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives. New York : Wiley*, 2004.

[62] M. Hof et A. Zwinderman, « Methods for analyzing data from probabilistic linkage strategies based on partially identifying variables », in : *Statistics in medicine* 31 (2012), 4231–4242.

[63] P Hougaard, « Fundamentals of survival data », eng, in : *Biometrics* 55.*1* (1999), p. 13, issn : 0006-341X.

[64] K. Imai et D. A Van Dyk, « Causal inference with general treatment regimes : Generalizing the propensity score », in : *Journal of the American Statistical Association* 99.*467* (2004), p. 854-866.

[65] M. Jamshidian et R. I. Jennrich, « Standard errors for EM estimation », in : *Journal of the Royal Statistical Society, Series B.* 62 (2000), p. 257-270.

[66] M. A. Jaro, « Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida », in : *Journal of the American Statistical Association* 84.*406* (1989), p. 414-420.

[67] R. I. Jennrich et M. D. Schluchter, « Unbalanced repeated-measures models with structured covariance matrices », in : *Biometrics* (1986), p. 805-820.

[68] J.D. Kalbfleisch et R.L. Prentice, *The Statistical Analysis of Failure Time Data*, New York : Wiley, 2002.

[69] L. Kantorovich, « On the transfer of masses », in : *Doklady Akademii Nauk SSSR* 37 (1942), p. 7-8.

[70] N. Keiding, P. K. Andersen et J. P. Klein, « The role of frailty models and Accelerated Failure Time models in describing heterogeneity due to omitted covariates », in : *Statistics in Medicine* 16.*2* (1997), p. 215-224.

[71] G. Kim et R. Chambers, « Regression analysis under incomplete linkage », in : *Computational Statistics & Data Analysis* 56.*9* (2012), p. 2756 -2770.

[72] J. P. Klein, C. Pelz et M. Zhang, « Modeling Random Effects for Censored Data by a Multivariate Normal Regression Model », in : *Biometrics* 55.*2* (mai 2004), p. 497-506, issn : 0006-341X.

[73] P. Lahiri et M. D. Larsen, « Regression Analysis with Linked Data », in : *Journal of the American Statistical Association* 100.*469* (2005), p. 222-230.

[74] N. M Laird et J. H. Ware, « Random-effects models for longitudinal data », in : *Biometrics* (1982), p. 963-974.

[75] P. Lambert et al., « Parametric accelerated failure time models with random effects and an application to kidney transplant survival », in : *Statistics in Medicine* 23.*20* (2004), p. 3177-3192.

[76] M. D. Larsen et D. B. Rubin, « Iterative Automated Record Linkage Using Mixture Models », in : *Journal of the American Statistical Association* 96.*453* (2001), p. 32-41.

[77] H. Li et al., « Heterogeneous Transfer Learning via Deep Matrix completion with Adversarial Kernel Embedding », en, in : *Proceedings of the AAAI Conference on Artificial Intelligence* 33.*01* (juill. 2019), Number : 01, p. 8602-8609, issn : 2374-3468.

[78] R.J. Little et D. Rubin, *Statistical Analysis with Missing Data*, Wiley, NY, 1987.

[79] H. P. Lopuhaä, « On the relation between S-estimators and M-estimators of multivariate location and covariance », in : *The Annals of Statistics* (1989), p. 1662-1683.

[80] Z. J. I. Lu, J. Sykes et M. Pintilie, « Understanding Dependent Competing Risks : A Simulation Study to Illustrate the Relationship Between Cause-Specific Hazard and Marginal Hazard », in : *Journal of Statistical Science and Application* 4.*03-04* (2016), p. 96-107.

[81] J. K. Lunceford, « Stratification and Weighting via the Propensity Score in Estimation of Causal Treatment Effects : A Comparative Study », English, in : *Statistics in Medicine* 36.*14* (juin 2017), p. 2320, issn : 1097-0258.

[82] J. K. Lunceford et M. Davidian, « Stratification and Weighting via the Propensity Score in Estimation of Causal Treatment Effects : A Comparative Study », English, in : *Statistics in Medicine* 23.*19* (oct. 2004), p. 2937-2960, issn : 1097-0258.

[83] A.-A. Mamun, R. Aseltine et S. Rajasekaran, « Efficient Record Linkage Algorithms Using Complete Linkage Clustering », in : *PloS one* 11 (avr. 2016), e0154446.

[84] X.L. Meng et D. Rubin, « Maximum Likelihood Estimation via the ECM Algorithm : A General Framework », in : *Biometrika* 80 (1993), p. 267-278.

[85] G. Monge, « Mémoire sur la Théorie des Déblais et des Remblais », in : *Histoire de l'Académie royale des sciences de Paris* (1781), p. 666-704.

[86] K. L. Moore et al., « Causal inference in epidemiological studies with strong confounding », in : *Statistics in medicine* 31.*13* (2012), p. 1380-1404.

[87] J. Neter, E. S. Maynes et R. R. Ramanathan, « The Effect of Mismatching on the Measurement of Response Errors », in : *Journal of the American Statistical Association* 60.*312* (1965), p. 1005-1027.

[88] S. K. Ng et G. J. McLachlan, « An EM-based semi-parametric mixture model approach to the regression analysis of competing-risks data », in : *Statist. Med.* 22.*7* (2003), p. 1097-1111, issn : 1097-0258.

[89] S. Noboa et al., « Estimation of a potentially preventable fraction of venous thromboembolism : a community-based prospective study », in : *Journal of Thrombosis and Haemostasis* 4.*12* (2006), p. 2720-2722.

[90] C. J. O'Brien, K. Petersen-Schaefer et al., « Adjuvant radiotherapy following neck dissection and parotidectomy for metastatic malignant melanoma », in : *Head and Neck* 19.*7* (1997), p. 589-594.

[91] G. Peyré et M. Cuturi, « Computational optimal transport », in : *Foundations and Trends® in Machine Learning* 11.*5-6* (2019), p. 355-607.

[92] J. C. Pinheiro, C. Liu et Y. N. Wu, « Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution », in : *Journal of Computational and Graphical Statistics* 10.*2* (2001), p. 249-276.

[93] I. Redko et al., « CO-Optimal Transport », in : *Advances in Neural Information Processing Systems*, t. 33, Curran Associates, Inc., 2020, p. 17559-17570.

[94] I. Redko et al., « Optimal transport for multi-source domain adaptation under target shift », in : *The 22nd International Conference on artificial intelligence and statistics*, PMLR, 2019, p. 849-858.

[95] N. Reid, « A Conversation With Sir David Cox », in : *Statistical Science* 9 (1994), p. 439-455.

[96] G. Renner et M. Reuschenbach, « Critical review on data processing algorithms in non-target screening : challenges and opportunities to improve result comparability », in : *Analytical and Bioanalytical Chemistry* 415.*18* (2023), p. 4111-4123.

[97] P. J. Rousseeuw, « Multivariate estimation with high breakdown point », in : *Mathematical statistics and applications* 8.*283-297* (1985), p. 37.

[98] Peter Rousseeuw et Victor Yohai, « Robust regression by means of S-estimators », in : *Robust and Nonlinear Time Series Analysis : Proceedings of a Workshop Organized by the Sonderforschungsbereich 123 "Stochastische Mathematische Modelle", Heidelberg 1983*, Springer, 1984, p. 256-272.

[99] M. Sadinle, « Bayesian Estimation of Bipartite Matchings for Record Linkage », in : *Journal of the American Statistical Association* 112.*518* (2017), p. 600-612.

[100] A. Sayers et al., « Probabilistic record linkage », in : *International journal of epidemiology* 45 (2015), 954–964.

[101] A. Skrondal et S. Rabe-Hesketh, *Generalized Latent Variable Modeling*, New York : Chapman et Hall/CRC, 2004.

[102] R. C. Steorts, R. Hall et S. E. Fienberg, « A Bayesian Approach to Graphical Record Linkage and Deduplication », in : *Journal of the American Statistical Association* 111.*516* (2016), p. 1660-1672.

[103] A. Sánchez Brotons et al., « Pipelines and Systems for Threshold-Avoiding Quantification of LC–MS/MS Data », in : *Analytical Chemistry* 93.*32* (2021), PMID : 34355890, 11215–11224.

[104] A. Tancredi et B. Liseo, « A hierarchical Bayesian approach to record linkage and population size problems », in : *The Annals of Applied Statistics* 5.*2B* (2011), p. 1553 -1585.

[105] M. Thorpe et al., « A Transportation Lp̂ Distance for Signal Analysis », in : *Journal of mathematical imaging and vision* 59 (2017), p. 187-210.

[106] V. Titouan et al., « Co-optimal transport », in : *Advances in Neural Information Processing Systems* 33 (2020), p. 17559-17570.

[107] A. Tsiatis, « A nonidentifiability aspect of the problem of competing risks », eng, in : *Proceedings of the National Academy of Sciences of the United States of America* 72.*1* (1975), p. 20, issn : 0027-8424.

[108] S. Van Aelst et G. Willems, « Multivariate regression s-estimators for robust estimation and inference », in : *Statistica Sinica* (2005), p. 981-1001.

[109] S. Van Buuren, « Multiple imputation of discrete and continuous data by fully conditional specification », in : *Statistical methods in medical research* 16.*3* (2007), p. 219-242.

[110] V. N. Vapnik, *The Nature of Statistical Learning Theory*, New York, NY, USA : Springer, second edition, 2000, isbn : 0-387-94559-8.

[111] B. Vellas et al., « Long-term use of standardised ginkgo biloba extract for the prevention of Alzheimer's disease (GuidAge) : a randomised placebo-controlled trial », in : *Lancet Neurology* (2012).

[112] C. Villani, « Optimal transport, old and new », in : *Grundlehren des mathematischen Wissenschaften, Springer-Verlag* 338 (2009).

[113] L.-J. Wei, « The accelerated failure time model : a useful alternative to the Cox regression model in survival analysis », in : *Statistics in medicine* 11.*14-15* (1992), p. 1871-1879.

[114] E. J. Williamson, A. Forbes et I. R. White, « Variance Reduction in Randomised Trials by Inverse Probability Weighting Using the Propensity Score », English, in : *Statistics in Medicine* 33.*5* (fév. 2014), p. 721-737, issn : 1097-0258.

[115] W. E. Winkler, « Frequency-based Matching in the Fellegi-Sunter Model of Record Linkage », in : *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1989, p. 778-783.

[116] W. E. Winkler, « Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage », in : *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1988, p. 667-671.

[117] C. F. J. Wu, « On the Convergence Properties of the EM Algorithm », in : *The Annals of Statistics* 11.*1* (mars 1983), p. 95-103.

[118] H. Xu et al., « Incorporating conditional dependence in latent class models for probabilistic record linkage : Does it matter ? », in : *The Annals of Applied Statistics* 13 (sept. 2019), p. 1753-1790.

[119] Y. Yan et al., « Semi-Supervised Optimal Transport for Heterogeneous Domain Adaptation », in : *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, International Joint Conferences on Artificial Intelligence Organization, juill. 2018, p. 2969-2975.

[120] Y. Yao et al., « Heterogeneous Domain Adaptation via Soft Transfer Network », in : *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, New York, NY, USA : Association for Computing Machinery, oct. 2019, p. 1578-1586, isbn : 978-1-4503-6889-6.

[121] L.-C. Zhang et T. Tuoto, « Linkage-data linear regression », in : *Journal of the Royal Statistical Society : Series A (Statistics in Society)* 100 (2020), 222–230.

[122] L.-C. Zhang et T. Tuoto, « Linkage-data linear regression », in : *Journal of the Royal Statistical Society : Series A (Statistics in Society)* 184.*2* (2021), p. 522-547.

[123] Z. Zhang et al., « Causal inference with a quantitative exposure », in : *Statistical methods in medical research* 25.*1* (2016), p. 315-335.

[124] V. J. Zhu et al., « An Empiric Modification to the Probabilistic Record Linkage Algorithm Using Frequency-Based Weight Scaling », in : *Journal of the American Medical Informatics Association* 16.*5* (sept. 2009), p. 738-745.

[125] B. Zou et al., « On variance estimate for covariate adjustment by propensity score analysis », ENG, in : *Statistics in Medicine* 35.*20* (sept. 2016), p. 3537-3548, ISSN : 1097-0258.

Née le 23/05/1986
Nationalité : Française
+33 6 84 74 41 20
valerie.gares@inria.fr
http ://vgares.perso.math.cnrs.fr/
2 enfants

Parcours Académique

2014 **Doctorat de Mathématiques Appliquées**. Université Paul Sabatier, Toulouse, France.
2013 **Obtention de l'agrégation externe de Mathématiques**.
2010 **Master 2 Recherche Mathématiques Appliquées**, filière Probabilité et Statistique. Université Paul Sabatier.
2008 **Obtention du Capes externe de Mathématiques**.
2008 **Master 1 Mathématiques Fondamentales**. Université Paul Sabatier.

Thématiques de recherche

- **Analyse de données de survie**.
  - ✓ Statistiques et tests non-paramétriques de comparaison de groupes en survie.
  - ✓ Calcul Stochastique. Processus de comptage et martingales. Efficacité asymptotique de Pitman.
  - ✓ Modèles à risques compétitifs. Modèles de régression multivariés linéaires censurés. Algorithme EM.
- **Intégration de données multi-sources**.
  - ✓ Appariement statistique, apprentissage par transfert, adaptation de domaine hétérogène : algorithme basé sur la théorie du transport optimal
  - ✓ Chaînage des données. Modèles probabilistes. Prise en compte de l'erreur d'appariement dans les modèles statistiques.
- **Analyse robuste de données longitudinales balancées et non balancées**.
  - ✓ Modèle de régression linéaire mixte. S estimateur.
- **Analyse causale**.
  - ✓ Score de propension. Estimation de variance.
- **Analyse de données fonctionnelles avec des structures groupées**.
  - ✓ Modèle à classes latentes. ACP probabiliste.
- **Apprentissage statistique**.

Domaines d'applications :

- Recherche clinique : maladie d'Alzheimer, cancérologie.
- Recherche épidémiologique : inégalité sociale de santé, santé et environnement. Prématurité.
- Ingénierie des infrastructures de transport.

Expérience Professionnelle

depuis 09/2024	**Chaire Professeur Junior**. INRIA de Rennes.
2017 – 2024	**Maîtresse de conférences**. INSA de Rennes. Titulaire de la PEDR.
2016 – 2017	**Post doctorat**. Université Paul Sabatier. INSERM UMR 1295 CERPOP (équipe EQUITY), Toulouse.
2014 – 2015	**Post doctorat**. NHMRC Clinical Trial Centre. Université de Sydney, Australie.
2013 – 2014	**ATER** à temps complet. Université Jean Jaurès, Toulouse.
2010 – 2013	**Contrat doctoral**. INSERM UMR 1295 CERPOP (équipe Vieillissement et maladie d'Alzheimer), Toulouse.
03 – 07/2010	**Stage de Master 2**. LN Pharma, Toulouse (Start-up en recherche clinique et recherche épidémiologique).

Activités d'Enseignement

Responsabilités / Conseils

Au niveau du département mathématiques appliqués à l'INSA, j'ai été responsable :

- **de la 3$^{\text{ème}}$ année** de la spécialité MA de 2019 à 2024.
- des **contrats professionnels** en 5$^{\text{ème}}$ année de 2021 à 2024.
- du **Master Data science** à l'INSA en 2021-2022.
- du **séminaire entreprise** en 3$^{\text{ème}}$ année de la spécialité MA de 2017 à 2024.
- J'ai été **membre du conseil du département** de 2021 à 2024.

Au niveau de l'INSA :

- Participation à une **commission Inter INSA** pour repenser le modèle social en 2021.
- Participation chaque année aux **journées portes ouvertes** de l'INSA.
- Participation aux **entretiens** organisés par le groupe INSA pour recruter ses futurs élèves-ingénieurs (1$^{\text{ère}}$ et 3$^{\text{ème}}$ année) en 2022 et 2023.

Enseignements.

Année	Niveau	Intitulé	Type	Nature	Effectifs
Rennes 1					
2024	Master 2	Apprentissage statistique pour la biologie	MODE : environnement, BMC : bio cellulaire et BIOINFO, Rennes 1	CM (9h), TD (9h), 22.5heTD	80
2024	Master 2	Apprentissage statistique	Sciences éco, Rennes 1	CM (7.5h), 15heTD	80
INSA Rennes					
2019-20	2ème année	Analyse III	STPI (cycle préparatoire), INSA	TD (26h), 26heTD	20-30
2017-24	2ème année	Probabilités	STPI, INSA	TD (42h), 42heTD	20-30
2017-20	3ème année	Probabilités	Tronc commun, INSA	TD (20h), 20heTD	20-30
2017-24	3ème année	Analyse de données	MA (département Mathématiques appliquées), INSA	CM (10h), TP (16h), 31heTD	20-30
2017-24	3ème année	Statistique Inférentielle	MA, INSA	TD (16h), TP (4h), 20heTD	20-30
2017-24	4ème année	Apprentissage statistique	MA, INSA	CM (12h), TD (12h), TP (12h), 42heTD	20-30
2017-24	5ème année	Fiabilité	MA, INSA	CM (8h), TD (6h), TP (4h), 23heTD	20-30
2017-24	4ème année	Bureau d'étude (avec entreprise)	MA, INSA	TD (10h)	20-30
2017-24	4ème année MA	Projet interdépartement	MA, INSA	TD (8h)	20-30
2017-24	4ème année	Projet recherche	MA, INSA	TD (10h)	20-30
2017-24	3, 4 ou 5ème année	Stage	MA, INSA	TD (1h)	20-30
MPSE Rennes					
2018 et 2019	5ème année	Modèles à risques compétitifs	Master Santé publique, parcours Modélisation en pharmacologie clinique et épidémiologie (MPSE)	Cours (4h)	15
Université Le Mirail, Toulouse					
2017	L3	Statistique inférentielle	MIASHS (Mathématiques et informatique appliquées aux sciences humaines et sociales)	Cours (18h), TD (18h). 45heTD.	20-30
2017	L2	Statistique inférentielle	MIASHS	Cours (18h), TD (18h). 45heTD.	20-30
2014	L1	Statistique descriptive	Psychologie	Cours (40h), TD (40h). 100heTD.	20-30
2013	L1 MIASHS	Analyse	MIASHS	TD. 47heTD.	20-30
2013	L2	Statistique inférentielle	Psychologie	Cours (18h), TD (18h). 45heTD.	20-30
Université de Sydney, Australie					
2015	Master 2	Projet Capsone, projet de recherche clinique	Recherche Essai Clinique	En ligne	
2015	Master 2	Principes d'inférences statistiques	Recherche Essai Clinique	En ligne	

Enseignement dans d'autres établissements.

Encadrement de **projets méthodologiques** en lien avec mes travaux de recherche :

- en 3ème année du parcours ingénieur à l'**ENSAI** en 2023-2024.
- en MASTER 2 du parcours **MIASH** (Rennes 2) en 2023-2024.

Enseignement à l'échelle internationale.

- Encadrement d'un groupe d'étudiantes sur un projet en statistique appliquée dans le cadre du **workshop "Women in Sage"** au Burundi en septembre 2024. Lien.
- **Apprentissage statistique**, niveau Master et doctorat, à M'bour au Sénégal en septembre 2022. J'y ai aussi encadré un projet de recherche de 3 étudiantes portant sur l'estimateur de Kaplan-Meier. Lien.
- **Analyse de données de survie**, niveau Master et doctorat au Bangladesh, (Mathematical and Computer Models in Epidemiology, Ecology and Agronomy) en mai 2022. Cet enseignement s'est effectué à distance. Lien.

Activités d'Encadrement

Encadrement de stage de recherche.

- Encadrement du stage de Lucas Offroy (4ème année INSA). *Algorithmes pour l'adaptation de domaines hétérogènes et application sur des données en génomiques* avec Chloé Friguet (IRISA, Vannes). 2024.

- Encadrement du stage de Florine Kempf (4ème année ENSAI). *Modélisation des parcours de soins et évolutions du système de santé* avec Madison Giacofci (Rennes 2) et Nolwenn Le Meur (EHESP). 2024.

- Encadrement du stage de Jean Dufraîche dans le cadre du parcours recherche de l'INSA (5ème année) et de son stage de fin d'étude. *Harmonizing Music and Mathematics : Exploring Musical Scores with Topological Data Analysis* avec Madison Giacofci (Rennes 2) et Nicolas Klutchnikoff (Rennes 2). 2023-2024.

- Encadrement du stage de Habib Toure (4ème année ENSAI). *Modélisation prédictive du devenir d'enfants prématurés* avec Jean-François Dupuy (INSA Rennes) et Patrick Pladys (CIC). 2022.

- Encadrement du stage de Marion Jeamart (Master 2, Rouen). *Causal Statistical Matching with Optimal Transport*. 2021.

- Encadrement du stage de Faisal Jayousi (4ème année M2, Rennes 2). *Optimal transport for statistical matching*. 2021.

- Encadrement du stage de (INSA Rennes). *Etude sur l'impact environnemental des activités de recherche de l'INSA de Rennes* avec Jérémy Omer. 2021.

- Encadrement du stage de César Arnould dans le cadre du parcours recherche de l'INSA (4ème année, INSA). *Exposition à la pollution atmosphérique et comportement de l'enfant dans les cohortes mère-enfant PELAGIE et EDEN* avec Nathalie Costet (IRSET). 2021.

- Encadrement du stage de Anaïs Foucault (5ème année ENSAI) avec Nathalie Costet. 2019.

- Encadrement du stage de Huan Võ Thành dans le cadre du master franco-vietnamien. *Probabilist models for record linkage* avec Guillaume Chauvet (ENSAI). 2019.

- Encadrement des analyses statistiques des thèses de Abel Guillen (psychiatrie) et Mathilde Bérodier (médecine générale). 2016.

- Encadrement du stage de El Hassane Ouaalaya (étudiant de Master 2 IMAT). *Data Clustering*. INSERM UMR 1295 CERPOP. 2014.

Encadrement doctoral.

Je suis encadrante actuellement de deux thèses :

- la thèse de Vanessa Chezeu (Cotutelle INSA/Cameroun) depuis septembre 2023 dont le sujet porte sur *l'analyse de données de survie de données chaînées*. Cette thèse est co-encadrée avec Jean-François Dupuy (taux d'encadrement : 25%).

- la thèse de Ghanem Bahrini (thèse CIFRE INSA / SAFRAN) depuis septembre 2023 dont le sujet porte sur *l'analyse de survie en présence de risques concurrents avec prise en compte des données longitudinales*. Cette thèse est co-encadrée avec Jean-François Dupuy (taux d'encadrement : 25%).

J'ai été encadrante de différentes thèses :

- la thèse de Renan Bernard dont le sujet porte sur *le transport optimal et la causalité*. Cette thèse a été co-encadrée avec Nicolas Courty (Université Bretagne Sud) et Chloé Friguet (IUT Vannes) (taux d'encadrement : 33%). L'étudiant a choisi d'interrompre sa thèse.

- la thèse de Abdoulaye Koroko dont la thématique, nouvelle pour moi, porte sur *les Méthodes d'optimisation stables et parallèles pour l'apprentissage avec de grandes données*. Cette thèse a été co-encadrée avec Ani Anciau Sedrakian (IFP Energies nouvelles), Ibtihel Ben Gharbia (IFP Energies nouvelles), Mounir Haddou (INSA Rennes) et Quang Huy tran (IFP Energies nouvelles) (taux d'encadrement : 20%). Il a soutenu le 16 octobre 2023.

- la thèse CIFRE à EUROVIA de Aymane Gourimane dont le sujet porte sur *l'Amélioration de l'uni longitudinal des chaussées*. Cette thèse a été co-encadrée avec Fekri Meftah (INSA Rennes) et Jean-François Dupuy (INSA Rennes) (taux d'encadrement : 33%). Il a soutenu le 20 juin 2023. Elle résulte d'une collaboration avec le département Génie Civil à l'INSA.

- la thèse de Huan Võ Thành dont le sujet porte sur *les modèles de couplage d'enregistrements probabilistes et l'analyse de données couplées avec une application sur les données du SNDS*. Cette thèse a été co-encadrée avec Guillaume Chauvet (ENSAI Rennes), Stéphane Paquelet (IRT bcom) et André Happe (CIC, Rennes) (taux d'encadrement : 42,5%). Il a soutenu le 1er décembre 2022. Elle s'inscrit dans le cadre d'une collaboration avec l'EHESP.

Publications.

Les noms soulignés correspondent aux noms des doctorants que j'ai encadrés.

- Articles publiés

1. **Garès V.**, Hudson M., Manuguerra M. et Gebski V. *Effect of a correlated competing risk on marginal survival estimation in an accelerated failure time model.* Communications in Statistics - Simulation and Computation. pp. 1-23, 2024. Article
2. Koroko A., Anciaux-Sedrakian A., Ben Gharbia I., **Garès V.**, Haddou M. et Tran Q.H. *Analysis and Comparison of Two-Level KFAC Methods for Training Deep Neural Networks.* Optimization Methods and Software. pp 1-33, 2024. Article.
3. Lopuhaä H.P., **Garès V.** et Ruiz-Gazen A. *S-estimation in linear models with structured covariance matrices*. Annals of statistics. 51(6), pp. 2415-2439, 2023. Article.
4. Vo T.H., **Garès V.**, L-C. Zhang L-C., Happe A., Oger E., S. Paquelet S. et Chauvet G. *Cox regression with linked data*. Statistics in medecine. 43(2), pp. 296-314, 2023. Article.
5. Vo T.H., Chauvet G., Happe A., Oger E., Paquelet S. et **Garès V.** *Extending the Fellegi-Sunter record linkage model for mixed-type data with application to the French national health data system*. Computational Statistics and Data Analysis journal. 179, pp. 107656, 2022. Article.
6. Koroko A., Anciaux-Sedrakian A., Ben Gharbia I., **Garès V.**, Haddou M. et Tran Q.H. *Efficient Approximations of the Fisher Matrix in Neural Networks using Kronecker Product*

Singular Value Decomposition. ESAIM : Proceedings and Survey. 73, pp. 218-237, 2022. Article.

7. Guernec G., **Garès V.**, Omer J., Saint-Pierre P. et Savy N. *OTrecod : An R Package for Data Fusion using Optimal Transportation Theory*. 64(1), 33-56, R Journal. 2023. Article.
8. **Garès V.**, Hajage D. et Chauvet G. *Closed-form variance estimator for generalized propensity score*. Biometrical Journal. 2020. Article.
9. **Garès V.**, Omer, J. *Regularized optimal transport of covariates and outcomes in data recoding*. Journal of American Statistical Association, 117(537), pp. 320-333, 2020. Article.
10. **Garès V.**, Dimeglio C, Kosorok M. R., Guernec G., Fantin R., Lepage B. et Savy N. *On the use of optimal transportation theory to recode variables and application to database merging*. International Journal of Biostatistics, 16(1), 2019. Article.
11. Gebski V., **Garès V.**, Gibbs E. et Byth K. *Data maturity and follow-up in time-to-event analyses : How far in time to extend the Kaplan-Meier Plot*. International Journal of Epidemiology, 47(3), pp. 850-859, 2018. Article.
12. **Garès V.**, Andrieu S., Dupuy J.-F. et Savy N. *On Fleming-Harrington's test for late effects in prevention randomized controlled trials*. Journal of Statistical Theory and Practice, 11(3), pp. 418-435, 2017. Article.
13. **Garès V.**, Andrieu S., Dupuy J.-F. et Savy N. *An omnibus test for several hazard alternatives in prevention randomized controlled clinical trials*. Statistics in Medicine, 34(4), pp. 541-557, 2015. Article.
14. **Garès V.**, Andrieu S., Dupuy J.-F. et Savy N. *A comparison of the constant piecewise weighted logrank and Fleming-Harrington tests*. Electronic Journal of Statistics, 8(1), pp. 841-860, 2014. Article.

- Articles appliqués au médical. Je suis premier auteur de l'article numéro 5, en épidémiologie sociale.

1. Lequy E., Javad M., Vienneau D., de Hoogh K., Chen, J., Dupuy, J-F., **Garès V.**, Burte, E., Bouaziz, O., Le Tertre A., Wagner, V., Hertel O., Christensen J. H., Zhivin S., Siemiatycki J., Goldberg M., Zins M. et Jacquemin B.. *Influence of exposure assessment methods on associations between long-term exposure to outdoor fine particulate matter and risk of cancer in the French cohort Gazel*. Science of The Total Environment, 820, pp. 153098, 2022. Article.
2. Lequy E., Siemiatycki J., De Hoogh K., Vienneau D., Dupuy J.-F., **Garès V.**, Hertel O., Christensen J. H., Zhivin S., Goldberg M., Zins M. et Jacquemin B. *Contribution of Long-Term Exposure to Outdoor Black Carbon to the Carcinogenicity of Air Pollution : Evidence regarding Risk of Cancer in the Gazel Cohort*. Environmental Health Perspectives, 129(3), pp. 37005, 2021. Article

3. Guillen A., Curot J., Birmes P., Denuelle M., **Garès V.**, Taïb S., Valton L. et Yrondi A. *Suicidal Ideation and Traumatic Exposure Should Not Be Neglected in Epileptic Patients : A Multi-dimensional Comparison of the Psychiatric Profile of Patients Suffering from Epilepsy and Patients Suffering from Psychogenic Non-epileptic Seizures.* Frontiers in Psychiatry, section Mood and Anxiety Disorders, 10, pp. 303, 2019. Article.
4. Castagné R., **Garès V.**, Karimi M., Chadeau-Hyam M., Vineis P, Delpierre C. et Kelly-Irving M. for the Lifepath consortium. *Allostatic load and subsequent all-cause mortality : which biological markers drive the relationship ? Findings from a UK birth cohort.* Europeen Journal of Epidemiology, 33(5), pp. 441-458, 2018. Article.
5. **Garès V.**, Panico L., Castagné R., Delpierre C. et Kelly-Irving M. *The role of the early social environment on Epstein Barr virus infection : a prospective observational design using the Millennium Cohort Study.* Epidemiology and infections, 145(16), pp. 3405-3412, 2017. Article.
6. Garsed D. W., Alsop K., Fereday S., Emmanuel C., Kennedy C., Etemadmoghadam D., Gao B., Gebski V., **Garès V.**, Christie E. L., Wouters M. C. A., Milne K., George J., Patch A.M., Li J., Arnau G. M., Semple T., Gadipally S., Chiew Y.E, Hendley J., Mikeska T., Zapparoli G. V., Amarasinghe K., Grimmond S. M., Hung J., Stewart C., Sharma R., Allen P., Millstein J., Rambau P. F.K, The Australian Ovarian Cancer Study [...] Bowtell D. D. L. et DeFazio A. *Homologous Recombination DNA Repair Pathway Disruption and Retinoblastoma Protein Loss are Associated with Exceptional Survival in High-Grade Serous Ovarian Cancer.* Clinical Cancer Research, 24(3), pp. 569-580, 2017. Article.
7. Coley N.*, Gallini A.*, **Garès V.**, Gardette V., Andrieu, S. et the ICTUS/DSA group. *A longitudinal study of transitions between informal and formal care in Alzheimer's disease using multistate models in European ICTUS cohort.* Journal of the American Medical Directors Association, 16(12), pp. 1104.e1-7, 2015. Article.
8. Soto M., Andrieu S., **Garès V.**, Gillette-Guyonnet S., Cantet C., Vellas B et Nourhashémi F. *Living alone with Alzheimer's disease and the risk of adverse outcomes. Results from the PLASA study.* Journal of the American Geriatrics Society, 63(4), pp. 651-8, 2014. Article.

- Logiciels. **Packages** R :
  - ✓ Package R (2020) "OTRecord"
    Titre : Optimal transportation for data recoding.
    Auteurs : Guernec G., Navarro P., Garès V.
    (Package également disponible sur Julia)
  - ✓ Package R (2019) "bnc" :
    Titre : Bivariate normal censored linear model for competing risks.
    Auteurs : Hudson H. M., Manuguerra M., and Garès V.

Communications

- Avec contribution

1. Chezeu V. (or.), Dupuy J.F. et **Garès V.** *Statistical analysis of matched survival data in national health databases.* Journées de Biostatistique, Paris, 2024.
2. Chezeu V. (or.), Dupuy J.F. et **Garès V.** *Statistical analysis of matched survival data in national health databases.* 55ème journées de la Société Française de Statistique, Bordeaux, 2024.
3. Dufraiche, J. (or.), **Garès V.**, Giacofci M. et Klutchnikoff N.. Exploring *Optimal Transport in Jazz Music Analysis. Application to the Real Book.* 55ème journées de la Société Française de Statistique, Bordeaux, 2024.
4. **Garès V.** (or.), Lopuhaä R. et Ruiz Gazen A. *Robust estimation in linear mixed effects models.* 55ème journées de la Société Française de Statistique, Bordeaux, 2024.
5. Giacofci M. (or.), **Garès V.** et Monbet V. *Mixture of functional principal component regression.* 54ème journées de la Société Française de Statistique, Bruxelles, 2023.
6. **Garès V.** (or.), Courty N., Friguet C., Gavra I., Klutchnikoff N. et Navaro P. *Optimal transport for data integration.* 54ème journées de la Société Française de Statistique, Bruxelles, 2023.
7. Vo T.H. (or.), **Garès V.**, Happe A., Oger E., Paquelet S. et Chauvet G. *Cox regression with linked data.* The 43rd Annual Conference of the International Society for Clinical Biostatistics (ISCB). Newcastle upon Tyne, UK, 21-25 août 2022.
8. Lavenu A. (or.), Murris J., Mareau A., Rouzé T., Fromont M., **Garès V.** et Katsahian S. *Comparaisons de méthodes pour données de survie en grande dimension sur de petits échantillons : optimisation des hyperparamètres et validation.* Conférence Intelligence artificielle et santé : approches interdisciplinaires, Nantes, 29 juin au 1er juillet 2022.
9. **Garès V.** (or.), Hajage D. et Chauvet G. *Variance estimators for weighted and stratified linear dose-response function estimators using generalized propensity score.* 53ème journées de la Société Française de Statistique, Lyon, 16 juin 2022.
10. Vo T.H. (or.), **Garès V.**, Happe A., Oger E., Paquelet S. et Chauvet G. *Cox regression with linked data.* 53ème journées de la Société Française de Statistique, Lyon, 16 juin 2022.
11. Lavenu A. (or.), Murris J., Mareau 1., Rouzé T., Fromont M., **Garès V.** et Katsahian S. *Comparaisons de méthodes pour données de survie en grande dimension sur de petits échantillons : optimisation des hyperparamètres et validation.* 53ème journées de la Société Française de Statistique, Lyon, 16 juin 2022.
12. Friguet C. (or.), Jeamart M., Bernard R., Courty N. et **Garès V.** *JDCOT : an Algorithm for Transfer Learning in Incomparable Domains using Optimal Transport.* 53ème journées de la Société Française de Statistique, Lyon, 14 juin 2022.

13. **Garès V.** (or.), Hajage D. et Chauvet G. *Variance estimators for weighted and stratified linear dose-response function estimators using generalized propensity score.* ISCB 2021, Lyon, 20 juillet 2021.

14. Vo T.H. (or.), Chauvet G. , Happe A., Oger E., Paquelet S. et **Garès V.** *An extension of Fellegi-Sunter record linkage model for mixed-type data with application to SNDS.* The 42rd Annual Conference of the International Society for Clinical Biostatistics (ISCB). Lyon, France, 18-22 Juillet 2021.

15. Vo T.H. (or.), Chauvet G. , Happe A., Oger E., Paquelet S. et **Garès V.** *An extension of Fellegi-Sunter record linkage model for mixed-type data with application to SNDS.* 52ème Journées de Statistique. Nice, France, 7-11 Juin 2021.

16. Hudson M. (or.), **Garès V.**, Manuguerra M. et Gebski V. *Correlated bivariate Normal competing risks – simulation findings in an ill-posed problem.* Joint International Society for Clinical Biostatistics and Australian Statistical Conference, Melbourne, 30 août 2018.

17. **Garès V.** (or.), Omer J., Guernec G. et Savy N. *On the use of optimal transportation theory to merge databases.* 50ème journées de la Société Française de Statistique, EDF Lab Paris Saclay, 31 mai 2018.

18. Delpierre C. (or.), **Garès V.**, Panico L., Castagné R. et Kelly-Irving M. *What is the role of the early social environment on Epstein-Barr Virus infection ? A prospective observational design using the Millennium Cohort Study.* Society for Longitudinal and Life Course Studies, Stirling, Scotland UK, octobre 2017.

19. **Garès V.**, Delpierre C., Kelly M. et Savy N. *BMI trajectories of children from the Millenium Cohort Study.* Journées du GDR statistiques et santé, Lyon, 28 juin 2016.

20. **Garès V.** (or.), Hudson M. et Gebski V. *Estimating correlation between competing risks.* Biometrics by the Harbour, Hobart, Tasmanie, Autralie, 3 décembre 2015.

21. Gibbs E. (or.), Gebski V., **Garès V.** et Byth K. *Data maturity and follow-up in time-to-event analyses : How far in time to extend the Kaplan-Meier Plot ?* ACTA 2015, International Clinical Trials Symposium, Sydney, Australie, 9 Octobre 2015.

22. **Garès V.** (or.), Hudson M. et Gebski V. *Estimating correlation between competing risks.* ACTA 2015, International Clinical Trials Symposium, Sydney, Autralie, 9 Octobre 2015.

23. Savy N. (or.), **Garès V.**, Andrieu S. et Dupuy J.-F. *On the use of Fleming and Harrington's test to detect late effects in clinical trials.* 8th International Conference of the ERCIM Working Group on Computational and Methodological Statistics, Londres, UK, 2015.

24. Coley N. (or.), **Garès V.**, Gallini A., Gardette V., Vellas B., Grand A., Andrieu S. et Ictus study group. *Evolution de la prise en charge des sujets diagnostiqués Alzheimer à 2 ans de suivi en Europe.* 34ème Journées Annuelles de la Société Française de Gériatrie et Gérontologie, Paris, France, 25-27 novembre 2014.

25. **Garès V.** (or.), Andrieu S., Dupuy J.-F. et Savy N. *Un nouveau test pour l'analyse de données de prévention en recherche clinique.* 45ème journées de la Société Française de Statistique, Toulouse, France, 27 mai 2013.
26. **Garès V.**, Andrieu S., Dupuy J.-F. et Savy N. *On the use of Fleming and Harrington test to detect Late Survival Difference in clinical trials.* Journées du GDR statistiques et santé, Rennes, 21 septembre 2012.
27. **Garès V.** (or.), Andrieu S., Dupuy J.-F. et Savy N. *On the use of Fleming-Harrington ?s test in prevention trials. Statistical Models and Methods for Reliability and Survival Analysis and Their Validation*, Bordeaux, France, 4 juillet 2012.
28. **Garès V.** (or.), Andrieu S., Dupuy J.-F. et Savy N. *Utilisation du test de Fleming-Harrington dans les essais de prévention.* 44ème journées de la Société Française de Statistique, Bruxelles, Belgique, 24 mai 2012.

- Invitations
  1. **Garès V.** (or.), Chauvet G., Chezeu V., Dupuy, J.F., Happe A., Oger E., Paquelet S et Vo T.H. *Statistical analysis of matched survival data in national health databases.* CMStatistics, décembre 2024.
  2. **Garès V.** (or.), Chauvet G., Chezeu V., Dupuy, J.F., Happe A., Oger E., Paquelet S et Vo T.H. *Statistical analysis of matched survival data in national health databases.* Colloque international de statistique et économétrie CISEM, juin 2024 (visio).
  3. **Garès V.** (or.), Courty N., Friguet C., Gavra I., Klutchnikoff N. et Navaro P. *Optimal transport for data integration.* journées de Statistique Mathématique et Data Science, Mammanet, Tuinisie, 12 novembre 2023 (en visio).
  4. **Garès V.** (or.), Jeamart M., Bernard R., Courty N. et Friguet C. *JDCOT : an Algorithm for Transfer Learning in Incomparable Domains using Optimal Transport.* Journées MAS 2022, Rouen, 29 août 2022.
  5. **Garès V.** (or.), Omer J., Guernec G. et Savy N. *Regularized optimal transport of covariates and outcomes in data recoding.* CMStatistics 2019, Londres, 14 décembre 2019.
  6. **Garès V.** (or.), Dimeglio C., Guernec G., Fantin R., Lepage B., Kosorok M.R. et Savy N. *On the use of optimal transportation theory to merge databases.* Conference on Statistics and Health , Toulouse, France, 12 janvier 2018.
  7. **Garès V.**, Andrieu S., Dupuy J.-F. et Savy N. *On the use of Fleming and Harrington test to detect Late Survival Difference in clinical trials.* Workshop "Analyse de Données Longitudinales de Cancer : Modèles de Markov Cachés" , Toulouse, 19 octobre 2012.

- Posters

1. Guernec, G, **Garès V.**, Navaro, P. Omer, J., Saint-Pierre P. et Savy, N. *OTrecod : An package for data integration using Optimal Transportation theory.* Poster. useR ! 2019, Toulouse, 9-12 juillet 2019.
2. **Garès V.**, Andrieu S., Dupuy J.-F. et Savy N. *Améliorer la performance des outils statistiques des essais de prévention de la maladie d'Alzheimer.* Poster. Association France Alzheimer et Fondation de France, Paris, 3 juin 2014.

- Séminaires : LMBA, Vannes (31 janvier 2025), Stat-Optim IMT (Toulouse, 5 mars 2024), Visite de l'INSMI à l'IRMAR (27 octobre 2023), "quelques questions d'éthique liées aux algorithmes d'intelligence artificielle", Séminaire éthique de l'IRMAR (16 juin 2023), Obelix, Vannes (17 juin 2019), Toulouse School of economics, équipe MADS, Toulouse (18 avril 2019), IRSET, Rennes (25 septembre 2018), Journée de l'IRMAR, Rennes (10 octobre 2017), séminaire de statistique, Rennes (15 septembre 2017), Séminaire scientifique INSERM UMR 1295 CERPOP équipe EQUITY, Toulouse (12 juillet 2016 et 8 mars 2016), séminaire Biostatistiques, NHMRC, Sydney (17 septembre 2015 et 19 mars 2015), séminaire scientifique INSERM UMR 1295 CERPOP, Toulouse (13 décembre 2013), Association France Alzheimer et Fondation de France, Paris (20 novembre 2012), Journées de l'inauguration de l'UMR 1027 Inserm, Toulouse (15 mars 2012).

- Groupes de travail : Groupe de Travail "Chaînage", EHESP, Rennes (11 septembre 2018), Groupe de travail "Small area estimation" (modèle mixte dans le cadre de petits domaines, 2016), Groupe de travail Biostatistiques INSERM UMR 1295 CERPOP, Toulouse (27 janvier 2017 et 12 décembre 2012).

Vie Scientifique

Responsabilités collectives et tâches d'intérêt général.

Au niveau de l'IRMAR :

- membre du **conseil de l'IRMAR** depuis janvier 2022.
- membre de la **commission IRMAR ECO** depuis 2019. Lien de la page web.
- membre du **comité parité** de 2018 à 2022. Lien de la page web.
- membre de l'**équipe d'organisation du séminaire Movi** (Modélisation Mathématique pour les Sciences du Vivant) depuis 2023.
  - ✓ Séminaire mensuel autour de la modélisation mathématique pour les sciences du vivant.
  - ✓ Lien de la page web.

Au niveau de l'INSA :

- membre du **Comité social d'administration** (CSA) et suppléante de la formation spécialisée du CSA en 2023-2024.
- **élue membre du conseil scientifique** de l'INSA de 2021 à 2024.
- membre du **conseil composante de l'IRMAR** à l'INSA de 2018 à 2024.
- responsable du séminaire statistique à l'INSA de Rennes (depuis 2017).

Au niveau national :

- membre de la **cellule communication de la SFDS** de 2018 à 2023.
- membre du **GDR Statistique et santé** de 2019 à 2023.
  ✓ organisation des journées de biostatistiques annuelles.
  ✓ Lien de la page web.
- membre du **réseau Thématique Math Bio Santé**, créé par l'INSMI (bureau et conseil scientifique) depuis 2024.

Révisions d'articles.

- International Journal of Biostatistics (1 en 2021).
- Journal of Biopharmaceutical statistics (1 en 2019).
- Journal de la Société Française de Statistique (1 en 2022).
- International Journal of Statistics in Medical Research (1 en 2023).
- BMC Medical Research Methodology (1 en 2024).

Participation à des jurys.

- Examinatrice de la thèse de Lucas Anzelin. *Prédictions du parcours de soins et du développement des enfants vulnérables.* 1 décembre 2023.

- Membre du jury de la thèse de Aymane Gourimate. *Amélioration de l'uni longitudinal des chaussées.* 20 juin 2023.

- Membre du jury de la thèse de Thanh Huan Vo. *Couplage d'enregistrements et analyse des données couplées avec application dans le système national français des données de santé.* 1 décembre 2022.

- Examinatrice de la thèse de Paul Frelon. *Application du Transport Optimal à l'analyse de données de cytométrie.* 28 mars 2023.

- Examinatrice de la thèse de Bilel Bousselmi. *Problèmes d'estimation dans des modèles de régression non paramétrique et de Poisson.* 6 décembre 2021.

- Examinatrice de la thèse de Van Trinh Nguyen. *Modèles de régression pour données de comptage zéro-inflatées en présence de censure. Applications en économie de la santé*. 16 mars 2020.

Membre de **2 comités de suivi individuel** de thèse (un en tant que membre extérieur au domaine et un dans le domaine de la statistique appliquée à la recherche clinique).

Comités de sélection/jurys.

Membre de comités de sélection.

- PRAG à l'INSA en 2022.
- maître de conférences à l'INSA de Fès au Maroc en 2018.
- maître de conférences à l'université Jean Jaurès à Toulouse en 2022.
- maître de conférences à l'INSA de Rennes en 2023.
- maître de conférences à l'Université Rennes 2 en 2023.
- enseignant chercheur à l' EHESP en 2023.
- maître de conférences à l'INSA de Lyon en 2024.

Représentante d'une école d'ingénieurs pour un jury VAE à l'ENSAI en 2019.

Groupes de travail.

- Responsable du groupe de travail "**Biostatistiques**" de l'UMR CERPOP de l'Inserm à Toulouse en 2016 et 2017.
- Mise en place de **deux groupes de travail en collaboration avec l'EHESP, l'Agrocampus et l'ENSAI** dans le cadre de problématiques liées aux données du SNIIRAM (assurance maladie) en 2018 avec Guillaume Chauvet (ENSAI) et Marie Pierre Etienne (Agrocampus) :
  - Chaînage de bases de données,
  - Données de Survie, Score de Propension (Responsable de ce groupe).

- Membre du groupe de travail "**Small area estimation**" à Toulouse School of Economics en 2017.

Comités d'organisation / scientifique.

Comités d'organisation.

- Workshop MoVi le 31 mai 2023 à l'Agrocampus à Rennes.
- MathsC2+ à l'INSA de Rennes en 2022, 2023 et 2024.
- Filles, Maths et Informatique : une équation lumineuse en 2021 à l'INSA, en 2022 à Rennes

2, en 2023 à l'ENSAI, en 2024 à Rennes 1 et en 2025 à l'INSA de Rennes.
- Colloque Francophone International sur l'Enseignement de la Statistique (CFIES) les 24 et 25 novembre 2022 à l'ENSAI Rennes.
- Journées du GDR statistique et santé les 17 et 18 novembre 2022 à l'INSA de Rennes. Responsable de l'organisation.
- Conférence "IA et santé" à Nantes du 29 juin au 1er juillet 2022. Je faisais partie également du comité scientifique pour représenter le GDR statistique et santé.
- "School in machine learning and artificial intelligence. Part 2 :Entropies, divergences and distances and applications in machine learning" du 20 au 24 juin 2022 à Rennes 2. Je faisais partie également du comité scientifique.
- Séminaire "Méthodes de l'approche causale en épidémiologie" organisé par l'IRSET, l'EHESP, l'INSA, l'IRMAR et l'Université de Rennes 1 du 6 au 8 juillet 2021. Je faisais partie également du comité scientifique.
- Colloque Jeunes Probabilistes et Statisticiens du 25 au 29 octobre 2021 et du 1 au 6 octobre 2023 à Oléron.
- JSTAR 2019 (Journées de Statistique de Rennes) à l'INSA de Rennes. Je faisais partie également du comité scientifique.
- Workshop "Conference on Statistics and Health" à Toulouse les 11 et 12 janvier 2018.

Comités scientifique.

- JSTAR 2024 à Rennes 2.

Projets.

- Allocation d'installation scientifique par Rennes Métropole (10 000 euros) en 2019.
- Projet Exploratoire de Premier Soutien (PEPS) "Jeune chercheuse, jeune chercheur" en 2019 (3500 euros).
- Membre des Partenariats Hubert Curien (PHC) en 2019.
- Membre du projet BIDASA financé dans le cadre des Défis MASTODONS en 2017. Lien de la page web.
- Membre du projet "Clinical trials advances for better health outcomes" par "The National Health and Medical Research council (NHMRC) en 2014.

Vulgarisation scientifique.

- **Article de vulgarisation** de mes travaux de recherche au service de la santé dans la revue scientifique de l'INSA de Rennes (INSIDE LABS, la recherche à l'INSA Rennes) en 2023.

- Intervention dans les classes sur la planche de Galton à l'école Jules Isaac (Rennes) le 19 et 26 juin 2023 avec Anne Cuzol (UBS, Vannes) (deux demi-journées) et à l'école Louise Michel (Rennes) avec Madison Giacofci (une demi-journée).

- Vulgarisation scientifique au collègue Didier Daurat et au lycée Bagatelle à Saint-Gaudens en février 2022.

COLLEGE MATHS, TELECOMS
DOCTORAL INFORMATIQUE, SIGNAL
BRETAGNE SYSTEMES, ELECTRONIQUE

Université de Rennes